www.sapientia.uji.es | 163

Ejercicios y casos prácticos con datos de corte transversal para la iniciación a la econometría

Jordi Ripollés Piqueras
Inmaculada Martínez Zarzoso
Maite Alguacil Marí



Colección «Sapientia», núm. 163

EJERCICIOS Y CASOS PRÁCTICOS CON DATOS DE CORTE TRANSVERSAL PARA LA INICIACIÓN A LA ECONOMETRÍA

Jordi Ripollés Piqueras
Inmaculada Martínez Zarzoso
Maite Alguacil Marí

Departamento de economía

■ Códigos de asignatura: AE1021 / FC1021 / EC1021
Fundamentos de Econometría



Edita: Publicacions de la Universitat Jaume I. Servei de Comunicació i Publicacions
Campus del Riu Sec. Edifici Rectorat i Serveis Centrals. 12071 Castelló de la Plana
http://www.tenda.uji.es e-mail: publicacions@uji.es

Colección Sapientia 163
www.sapientia.uji.es
Primera edición, 2020

ISBN: 978-84-17900-23-6
DOI: http://dx.doi.org/10.6035/Sapientia163



Publicacions de la Universitat Jaume I es miembro de la UNE, lo que garantiza la difusión y comercialización de sus publicaciones a nivel nacional e internacional. www.une.es.



Este libro, de contenido científico, ha estado evaluado por personas expertas externas a la Universitat Jaume I, mediante el método denominado revisión por iguales, doble ciego.

ÍNDICE GENERAL

Presentación

Ejercicios y casos prácticos con datos de corte transversal para la iniciación a la econometría está planteado para ser un manual complementario en la preparación de la asignatura Fundamentos de Econometría, en los grados de Administración de Empresas, Finanzas y Contabilidad, y Economía. Para acceder a las muestras de datos con las que trabajaremos en algunos ejercicios del manual, pulse aquí.

Cabe advertir que, en ningún caso, el presente material sustituye a la bibliografía básica, la cual es estrictamente necesaria seguir para preparar con garantías el examen final de la asignatura. Dicha bibliografía básica se compone de los siguientes manuales:

- Matilla, Mariano, Pedro Pascual y Basilio S. Carnero. 2013. *Econometría y predicción*. UNED: McGraw Hill.
- Wooldridge, Jeffrey M. 2010. *Introducción a la econometría. Un enfoque moderno*. 5.ª edición. México: Cengage Learning Editores.

ISBN: 978-84-17900-23-6

DOI: http://dx.doi.org/10.6035/Sapientia163

0. Introducción al análisis estadístico con Gretl

La asignatura Fundamentos de Econometría contiene sesiones de laboratorio impartidas en aulas de informática. En estas sesiones se introduce al alumnado en el análisis de regresión de variables económicas a través de la resolución de ejercicios y problemas con Microsoft Excel o el programa informático Gretl (Gnu Regression, Econometrics and Time-series Library). Este último programa, desarrollado por Allin Cottrell de la Universidad de Wake Forest, permite llevar a cabo análisis estadísticos y estimaciones de modelos econométricos. Gretl no solo presenta una interfaz visual muy intuitiva que permite realizar de forma sencilla multitud de análisis cuantitativos, sino que también contiene un conjunto de bases de datos de muestra procedentes de diversos manuales de econometría (Ramanathan 2002, Wooldridge 2010, Stock y Watson 2012, Verbeek 2008, entre otros).

Gretl es software libre y puede descargarse en http://gretl.sourceforge.net/.

A modo de resumen, en la siguiente tabla se presentan las instrucciones Gretl que emplearemos con mayor frecuencia en las sesiones de laboratorio de la asignatura:

Tabla 0.1. Resumen de instrucciones Gretl.

Descripción	Ruta
Cargar datos de muestra	*Archivo / Abrir archivo de datos / Archivo de muestra…*
Importar archivos externos de distintos formatos, como CSV (.csv), ASCII (.txt), Excel (.xls, .xlsx), Stata (.dta), entre otros	*Archivo / Abrir archivo de datos / Archivo de usuario…*
Indicar al software qué tipo de datos se van a utilizar: datos de sección cruzada, series de tiempo, o datos de panel	*Datos / Estructura de datos…*

ISBN: 978-84-17900-23-6
DOI: http://dx.doi.org/10.6035/Sapientia163

Descripción	Ruta
Obtener estadísticos principales de una variable aleatoria (media, mediana, mínimo, máximo, desviación típica, coeficiente de variación, coeficiente de asimetría y coeficiente de exceso de curtosis)	*Click derecho sobre el nombre de la variable / Estadísticos principales…*
Obtener la distribución de frecuencias de una variable	*Click derecho sobre el nombre de la variable / Distribución de frecuencias…*
Obtener la matriz de correlaciones entre dos o más variables	*Seleccionando dos o más variables (mientras se pulsa Ctrl)/ Click derecho sobre el nombre de las variables / Matriz de correlaciones*
Obtener la representación del diagrama de dispersión o gráfico X-Y	*Ver / Gráficos / Gráfico X-Y (scatter)*
Estimar un modelo por mínimos cuadrados ordinarios	*Modelo / Mínimos Cuadrados Ordinarios*

Para más información, el propio software ofrece una Guía del Usuario en el menú Ayuda de la barra de herramientas.

Práctica 0

EJERCICIO 0.1. El fichero «Data_Valencia_pisos.gdt» contiene información sobre una muestra aleatoria de 387 pisos en venta en Valencia, extraída de Nestoria (www.nestoria.es) el 15 de abril del 2018. En concreto, disponemos de datos sobre el precio de venta en miles de euros (*precio*), el tamaño de la vivienda expresado en metros cuadrados (*m2*), así como el número de dormitorios (*dormitorios*).

a) Indique qué estructura de datos presenta el archivo (datos de corte transversal, series temporales o datos de panel). ¿Por qué?
b) Represente e interprete la distribución de frecuencias de la variable *precio*.
c) Calcule e interprete los estadísticos descriptivos de las variables *precio* y *m2*.
d) Represente e interprete un diagrama de dispersión (gráfico X-Y) que muestre la relación entre las variables *precio* y *m2*.
e) Calcule e interprete la correlación entre las variables *precio*, *m2* y *dormitorios*.

1. El modelo de regresión simple

La relación $y = f(x)$ puede estudiarse a través de un modelo econométrico simple:

(regresando o variable dependiente) (regresor o variable independiente)

$$y=\beta_0+\beta_1 x+u$$

(error o perturbación aleatória: aquellos factores inobservables distintos de x que afectan a y)

Por un lado, el parámetro de la constante β_0 indica el valor que toma y, cuando $x = 0$. Por otro lado, el parámetro de la pendiente β_1 proporciona información sobre cuánto varía y ante cambios de x, cuando permanecen invariables otros factores que pueden influir sobre y (*ceteris paribus*).[1] Para que esto último sea así, es necesario que pueda asumirse la independencia de u ante cambios de x (*supuesto de media condicionada nula*):

$$\beta_1 = \left.\frac{\Delta y}{\Delta x}\right|_{\Delta u=0} \leftarrow \boxed{\text{Si } E(u|x)=E(u)=0}$$

Consecuentemente, la función de regresión poblacional (FRP) relaciona linealmente el promedio de y, E (y), para los distintos valores de x que presentan los individuos de una población: $E(y|x) = \beta_0 + \beta_1 x$.

1. El modelo es lineal en los parámetros β. Es decir, β_1 recoge el cambio de y ante un cambio unitario de x, independientemente del nivel de x.

ISBN: 978-84-17900-23-6
DOI: http://dx.doi.org/10.6035/Sapientia163

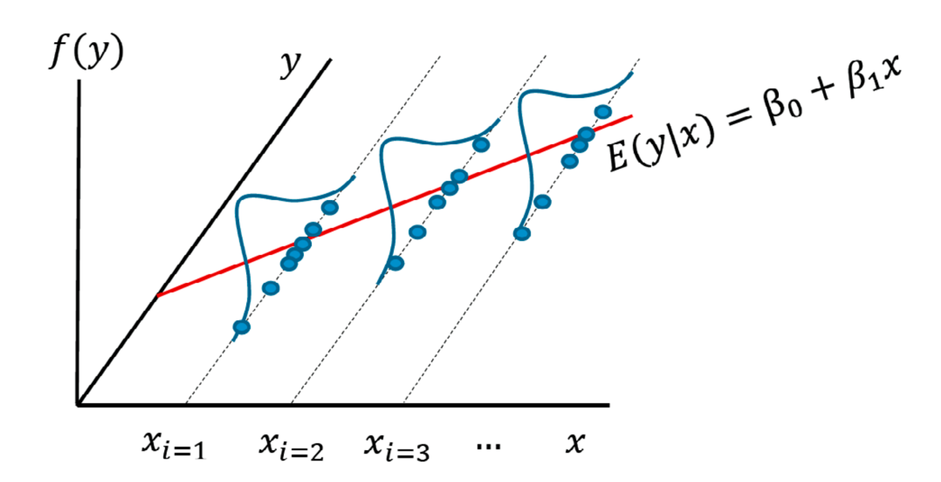


Figura 1.1. Función de regresión

Objetivo: evaluar el vínculo entre *y* y *x* mediante la estimación de los parámetros poblacionales β_0 y β_1 (fijos, pero desconocidos) a partir de un conjunto de observaciones de una muestra:

1. Utilizamos una muestra aleatoria de la población, $\{(y_i, x_i): i = 1,2,\ldots, n\}$.
2. Especificamos un modelo lineal en los parámetros β, para cada observación *i* de la muestra: $y_i = \beta_0 + \beta_1 x_i + u_i$.
3. Estimamos la función de regresión muestral (FRM) del modelo: $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$.

Métodos de estimación: A continuación, se presenta el método de los momentos (MM) y el método de mínimos cuadrados ordinarios (MCO). Dado que, en nuestro marco de trabajo, ambos métodos de estimación llegan al mismo resultado, en los ejercicios presentados a continuación se pedirá habitualmente estimar por MCO.[2]

2. Existen otros métodos de estimación, como el de máxima verosimilitud (MV), que consiste en seleccionar el valor de los parámetros que maximiza la probabilidad de obtener las observaciones muestrales. En los modelos lineales, la estimación por MV, la cual cumple las propiedades asintóticas deseables bajo condiciones más generales, también coincide con la obtenida por MCO para muestras grandes.

Método de los momentos (MM). Deseamos encontrar estimaciones de los parámetros poblacionales β_0 y β_1 que cumplan las siguientes dos restricciones:

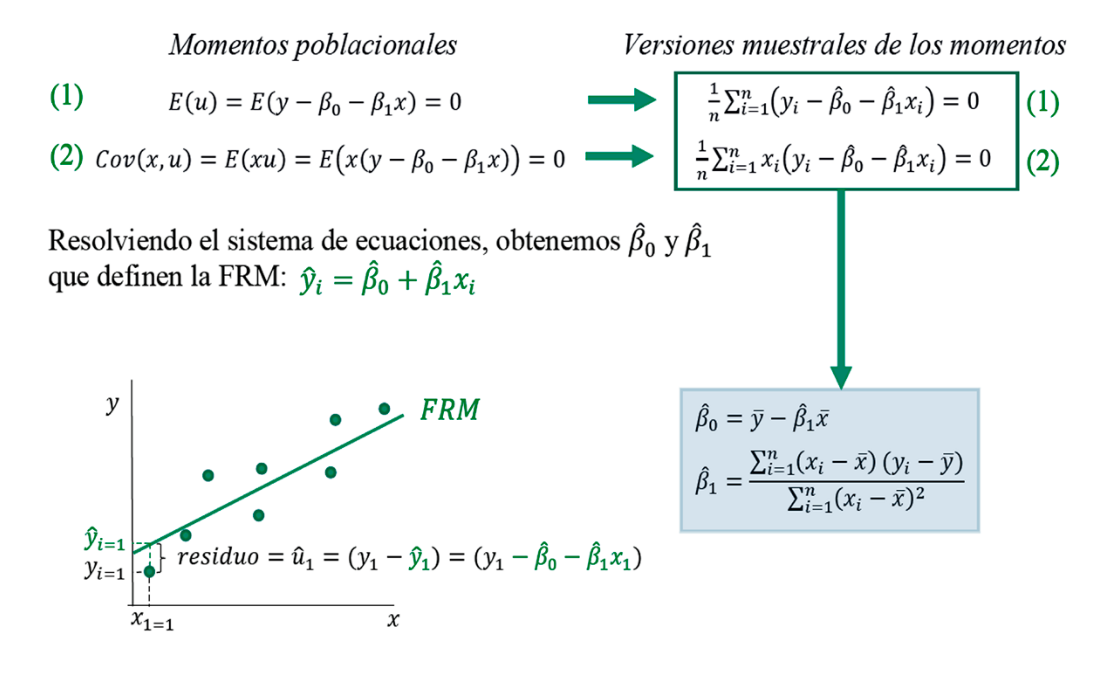


Figura 1.2 Valores observados (y_i) versus estimados ($\hat{y}_i$)

Nota: La representación de los valores observados en un gráfico de dispersión X-Y suele recibir el nombre de nube de puntos.

Al mismo resultado llegamos a través del *método de mínimos cuadrados ordinarios*(MCO). Este método estima $\hat{\beta}_0$ y $\hat{\beta}_1$, minimizando la suma de los residuos al cuadrado: $\min_{\hat{\beta}_0,\hat{\beta}_1} \sum_{i=1}^{n}(\hat{u}_i)^2 = \min_{\hat{\beta}_0,\hat{\beta}_1} \sum_{i=1}^{n}\left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i\right)^2$ donde las condiciones de primer orden (CPO) son, respectivamente, el análogo muestral de los momentos poblacionales (1) y (2):

$$\frac{\partial \sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2}{\partial \hat{\beta}_0} = 0 \rightarrow -2\sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0$$

$$\frac{\partial \sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2}{\partial \hat{\beta}_1} = 0 \rightarrow -2\sum_{i=1}^{n}x_i(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

Por tanto, cada observación $i = 1, 2, \ldots, n$ la variable y_i podrá expresarse como la suma de su predicción según la FRM ($\hat{y}_i$) y su residuo ($\hat{u}_i$): $y_i = \hat{y}_i + \hat{u}_i$.

Propiedades numéricas de la FRM:

1. (1) $\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0 \rightarrow \quad \sum_{i=1}^{n} \hat{u}_i = 0$
2. (2) $\frac{1}{n} x_i \sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0 \rightarrow \quad \sum_{i=1}^{n} x_i \hat{u}_i = 0$
3. $\bar{\hat{y}} = \bar{y}$ porque $\bar{y} = \bar{\hat{y}} + \bar{\hat{u}}$ (donde $\bar{\hat{u}} = 0$ por (1) y $\bar{\hat{y}} = \hat{\beta}_0 + \hat{\beta}_1 \bar{x}$)
4. La FRM se encuentra sobre el punto $(\bar{x}, \bar{y})$
5. Por (1) y (2), $\sum_{i=1}^{n} \hat{y}_i \hat{u}_i = 0$.

Bondad de ajuste: El coeficiente de determinación (R^2) nos permite conocer cómo de bien se ajusta nuestra FRM a la nube de puntos observados de y_i de la muestra. En términos más formales, R^2 indica qué proporción de la variabilidad muestral total exhibida por y viene explicada por x:

$R^2 = \frac{SEC}{STC} = 1 - \frac{SCE}{STC}$; $0 \leq R^2 \leq 1$ (cuando mayor R^2, mejor es el ajuste del modelo)

donde STC es la suma total de cuadrados, $\sum_{i=1}^{n}(y_i - \bar{y})^2$

SEC es la suma explicada de cuadrados, $\sum_{i=1}^{n}(\hat{y}_i - \bar{\hat{y}})^2$

SCE es la suma de los residuos al cuadrado, $\sum_{i=1}^{n} \hat{u}_i^2$

La siguiente tabla muestra las distintas formas funcionales, lineales en los parámetros, con las que podemos trabajar, así como sus principales características.

Tabla 1.1. Formas funcionales en el análisis de regresión.

Tipo	Modelo econométrico	Pendiente	Interpretación de la pendiente	Función de regresión
Nivel-nivel	$y = \beta_0 + \beta_1 x + u$	$\frac{\Delta y}{\Delta x} = \beta_1$	Por cada unidad que ↑x, y varía en β_1 unidades, con independencia del nivel de x.	$\beta_1 > 0$; $\beta_1 < 0$; $\beta_1 = 0$
Nivel-log o semielasticidad	$e^y = e^{\beta_0} x^{\beta_1} e^u$ $y = \beta_0 + \beta_1 log(x) + u$	$\frac{\Delta y}{\Delta log(x)} = \frac{\Delta y}{\Delta x/x} = \beta_1;$ $\frac{\Delta y}{\%\Delta x} = \frac{\beta_1}{100}$	Un ↑x en un 1% hace que y varíe en $(\beta_1/100)$ unidades. En otras palabras, β_1 es la semielasticidad de y respecto x.	$\beta_1 > 0$; $\beta_1 < 0$
Log-nivel o porcentaje constante	$y = e^{(\beta_0 + \beta_1 x + u)}$ $log(y) = \beta_0 + \beta_1 x + u$	$\frac{\Delta log(y)}{\Delta x} = \frac{\Delta y/y}{\Delta x} = \beta_1;$ $\frac{\%\Delta y}{\Delta x} = \beta_1 \cdot 100$	Por cada unidad adicional que ↑x, y varía en un $(\beta_1 \cdot 100)$%.	$\beta_1 > 0$; $\beta_1 < 0$
Log-log o elasticidad constante	$y = e^{\beta_0} x^{\beta_1} e^u$ $log(y) = \beta_0 + \beta_1 log(x) + u$	$\frac{\Delta log(y)}{\Delta log(x)} = \frac{\Delta y/y}{\Delta x/x} = \beta_1;$ $\frac{\%\Delta y}{\%\Delta x} = \beta_1$	Un ↑x en un 1% hace que y varíe en un β_1%. En otras palabras, β_1 es la elasticidad de y respecto x.	$\beta_1 > 1$; $\beta_1 < 0$; $0 < \beta_1 < 1$
Cuadrático[3]	$y = \beta_0 + \beta_1 x + \beta_2 x^2 + u$	$\frac{\Delta y}{\Delta x} = \beta_1 + 2\beta_2 x$	Por cada unidad adicional que ↑x, y varía en $(\beta_1 + 2\beta_2 x)$ unidades. Es decir, el efecto marginal de x sobre y depende del nivel de partida de x que se considere.	$\beta_1 > 0$, $\beta_2 < 0$; $\beta_1 < 0$, $\beta_2 > 0$

Nota: Siguiendo la misma nomenclatura que Wooldridge (2010), utilizaremos *log* para expresar logaritmo natural.[3]

Propiedades y aproximaciones empleadas:

$$log(x^y) = y \cdot log(x) \qquad log(x^\alpha \cdot y^\beta) = \alpha \cdot log(x) + \beta \cdot log(y)$$

$$log(e^\alpha) = \alpha \cdot log(e) = \alpha \qquad \Delta log(y) = log(y_1) - log(y_0) = log\left(\frac{y_1}{y_0}\right) = log\left(\frac{\Delta y + y_0}{y_0}\right) = log\left(\frac{\Delta y}{y} + 1\right) \approx \frac{\Delta y}{y}$$

3. En este caso, el valor de x que hace máxima o mínima la FRM $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x + \hat{\beta}_2 x^2$ viene dado por $\frac{\partial \hat{y}}{\partial x} = 0 \rightarrow \hat{\beta}_1 + 2\hat{\beta}_2 x = 0; \; x = \frac{\hat{\beta}_1}{2\hat{\beta}_2}$.

Práctica 1A

Ejercicio 1A.1

La siguiente tabla muestra, para un conjunto de hoteles de una localidad, el precio por noche de una habitación y el número medio de habitaciones ocupadas al día.

i	Precio (euros/noche)	Número de habitaciones ocupadas
1	35	150
2	100	20
3	90	50
4	115	10
5	70	100
6	60	130
7	50	180
8	80	100

Considere que deseamos explicar la relación entre la demanda hotelera y el precio a través del modelo:

$$Q_i = \beta_0 + \beta_1 P_i + u_i \quad \text{con} \quad i = 1, 2, \ldots, 8 \text{ hoteles}$$

donde Q representa el número de habitaciones ocupadas, y P es el precio por noche de la habitación. Con la información proporcionada, complete las siguientes tareas con la ayuda de Excel.

a) Represente e interprete un diagrama de dispersión (gráfico X-Y) que muestre la relación entre las dos variables (explicada y explicativa).

b) Utilice el procedimiento de mínimos cuadrados ordinarios (MCO) para estimar la función de regresión muestral del modelo planteado e interprete los valores estimados de la constante y de la pendiente.

c) De acuerdo con su estimación, ¿en cuánto se estima que varía la demanda hotelera si el precio de la habitación aumenta en 10 euros por noche?

d) Estime el número medio de habitaciones ocupadas para los niveles de precios observados en la muestra y, posteriormente, calcule los residuos. Además, compruebe si la suma de los residuos es aproximadamente 0.

e) ¿Qué proporción de la variabilidad muestral de la demanda es explicada por el precio?

f) En base a la función de regresión muestral obtenida, ahora prediga el número medio de habitaciones que se ocuparían si se fijase un precio de 75 euros por noche.

g) A partir de la función de regresión obtenida, calcule la elasticidad precio de la demanda para un precio de 75 euros por noche.[4]

h) Plantee un modelo que permita obtener directamente la elasticidad (constante) precio de la demanda, y explique qué parámetro en dicho modelo sería la elasticidad.

Ejercicio 1A.2

El fichero «Data_Gapminder_2010.gdt» contiene información extraída de Gapminder (*Free material from* www.gapminder.org) referente al 2010 para 247 países sobre distintas variables macroeconómicas, tales como el producto interior bruto (PIB), expresado en dólares per cápita, y la esperanza de vida, en años.

a) Represente el diagrama de dispersión (gráfico X-Y) con el PIB en el eje horizontal y la esperanza de vida en el eje vertical. ¿Qué tipo de relación diría que existe entre la esperanza de vida y el PIB?

b) En base al gráfico anterior, ahora plantee el modelo econométrico que mejor pueda describir la relación existente entre ambas variables y estímelo por MCO.

c) Interprete los parámetros estimados de la función de regresión muestral basada en el modelo planteado.

Ejercicio 1A.3

Busque en la red una muestra de datos de *corte transversal* para dos variables de índole económico-empresarial que crea que puedan estar relacionadas.[5] Utilizando dicha muestra de datos, realice las siguientes tareas:

a) Utilizando Excel, guarde en columnas las dos variables, nombrándolas y ordenándolas, junto a una variable índice *i*= 1, 2, ... , N para representar la dimensión de corte transversal (p. ej., personas, países, hogares, empresas, etc.). No olvide indicar la fuente de donde se han obtenidos los datos, así como el significado de cada variable y sus unidades de medida.

---

4. Recuerde que la elasticidad precio de la demanda se define como $\epsilon_p^d = \frac{\Delta Q}{\Delta P} \cdot \frac{P}{Q}$ donde $\frac{\Delta Q}{\Delta P} = \beta_1$ en la regresión.

5. Posibles fuentes de datos: Gapminder (www.gapminder.org), Goolzoom (www.goolzoom.es), Instituto Nacional de Estadística (www.ine.es), Eurostat (https://ec.europa.eu/eurostat/data/database), OECD (https://stats.oecd.org/), World Bank (https://data.worldbank.org/), UNCTADSTAT (https://unctadstat.unctad.org/), FAOSTAT (http://www.fao.org/faostat/en/#data), otros (https://www.economicsnetwork.ac.uk/links/sources, https://db.nomics.world/).

b) Utilice la teoría económica o el razonamiento lógico para justificar qué variable es la dependiente, *Y*, y cuál la explicativa, *X*.

c) Importe el archivo Excel a Gretl y represente e interprete la distribución de frecuencias y los estadísticos descriptivos de la variable dependiente que se pretende explicar (véase la solución del ejercicio 0.1).

d) Utilizando Excel o Gretl, calcule e interprete el coeficiente de correlación entre las variables seleccionadas (véase la solución del ejercicio 0.1).

e) Utilizando Excel o Gretl, represente e interprete un diagrama de dispersión (gráfico X-Y) que muestre la relación entre las dos variables (véase la solución del ejercicio 0.1).

f) En base al gráfico anterior, ahora plantee el modelo econométrico que mejor pueda describir la relación existente entre ambas variables y, posteriormente, estímelo por MCO utilizando Excel. *Muestre los cálculos efectuados en Excel en detalle y razone su respuesta* (véase la solución del ejercicio 1A.1).

g) ¿Qué proporción de la variabilidad muestral de la variable dependiente es explicada por el regresor? Use Excel para responder, y muestre los cálculos efectuados con detalle.

Ejercicio 1A.4

El fichero «Data_cons_inc.xlsx» contiene información procedente de Eurostat (Oficina Europea de Estadística, referencia: non-financial transactions, «nasq_10_nf_tr») sobre el consumo (consumo) y la renta disponible (rentad) en 2016, ambas expresadas en millones de euros, de 15 países europeos.

i Países	consumo (y)	rentad (x)
1. Alemania	1.674.394	1.970.801
2. Austria	186.225	213.596
3. Bélgica	216.574	241.024
4. Dinamarca	131.609	139.498
5. España	644.719	700.113
6. Finlandia	119.005	127.195
7. Francia	1.232.883	1.425.435
8. Grecia	121.737	114.009
9. Irlanda	90.847	94.739
10. Italia	1.022.411	1.137.017

i Países	consumo (*y*)	rentad (*x*)
11. Luxemburgo	16.037	20.071
12. Países Bajos	310.692	337.048
13. Portugal	121.335	128.768
14. Reino Unido	1.577.330	1.626.064
15. Suecia	205.911	235.318

a) Utilice la muestra aleatoria de tamaño *n*= 15 para estimar por MCO el siguiente modelo:

$$\text{consumo}_i = \beta_0 + \beta_1\, rentad_i + u_i \qquad \text{donde} \qquad i = 1, 2, \dots, 15.$$

Interprete los parámetros estimados de la constante y la pendiente. Según la estimación, ¿en cuánto se predice que variará el consumo si la renta disponible aumenta en un millón de euros?

b) De acuerdo con las estimaciones obtenidas, calcule el consumo predicho cuando la renta disponible es de 50.000 millones de euros.

c) En base a los resultados estimados, represente gráficamente y comente el comportamiento de las siguientes medidas en relación a la renta disponible:

- El consumo estimado, $\widehat{cons}$
- La propensión marginal a consumir estimada, $PMgC = \frac{\partial \widehat{cons}}{\partial inc}$
- La propensión media al consumo estimada, $PMEC = \frac{\widehat{cons}}{inc}$

d) Obtenga de Eurostat u otra fuente de información una segunda muestra de datos de corte transversal del mismo tamaño (n = 15) sobre consumo y renta. Indique si es posible validar los resultados obtenidos en los apartados anteriores.

Práctica 1B

Ejercicio 1B.1

Empleando una muestra aleatoria de 1.573 individuos españoles en el 2012 procedente de la European Social Survey (www.europeansocialsurvey.org), se ha obtenido el siguiente modelo estimado que relaciona el nivel de bienestar de los individuos con su edad:

$$log\ (happy) = 2{,}255\text{-}\ \beta_1\ \ 0.037\ \log\ (age)$$
$$n = 1573 \text{ y } R^2 = 0{,}002$$

donde *happy* es una variable que recoge la puntuación del 1 al 11 que responden los individuos encuestados a la pregunta: ¿cuánto de feliz es usted?; y *age* se refiere a los años de edad que tienen los individuos al responder la encuesta.

a) Interpretelos valores estimados de la constante y el coeficiente asociado a log (*age*).
b) Indique qué otras variables podrían influir en la felicidad de los individuos y explique si alguna/s de ellas podría/n estar correlacionadas con la edad. Si esto último pudiese ocurrir, ¿podríamos confiar en los resultados de la regresión simple del enunciado? ¿por qué?

Ejercicio 1B.2

El fichero «Data_salarios2014ESP.gdt» contiene información sobre los salarios percibidos (variable *salbase*, expresada en euros/mes) y los años de antigüedad en la empresa (*antig*) para el 2014 sobre una muestra de trabajadores que residen y trabajan en España. Dicha información ha sido extraída de la Encuesta de Estructura Salarial del INE.

a) Calcule la media y la desviación estándar tanto del salario como del número de años de antigüedad de los trabajadores de la muestra.
b) ¿Cuál es la proporción de individuos de la muestra que tienen menos de un año de antigüedad en la empresa ($antig < 1$)?, ¿cuál es el número máximo de antigüedad en la muestra?
c) Estime el siguiente modelo de regresión: $salbase = \beta_0 + \beta_1 \text{antig} + u$, y exponga los resultados de la función de regresión muestral. Según la estimación, interprete el término constante y la pendiente.
d) Según el modelo anteriormente estimado, ¿qué proporción de la variabilidad muestral exhibida por el salario viene explicada por la antigüedad?
e) Plantee y estime un modelo de regresión que permita predecir la variación porcentual salarial ante cada año adicional de antigüedad. Interprete el término constante e indique cuál sería el incremento porcentual estimado del salario ante un aumento de la antigüedad en 15 años.
f) Plantee y estime un modelo de regresión que permita obtener directamente la elasticidad de los salarios ante cambios en la antigüedad. Interprete el término constante, e indique cuál sería el incremento porcentual estimado del salario si se duplica la antigüedad.

Ejercicio 1B.3

La base de datos «Data_RD_scoreboard.gdt» (fuente: R&D Scoreboard de la Comisión Europea: http://iri.jrc.ec.europa.eu/scoreboard16.html) contiene información sobre las 2.500 empresas que más invierten en investigación y desarrollo en el mundo. Para la muestra de empresas mencionada, las variables *rd* y *sales* representan, respectivamente, los gastos en I+D y las ventas, ambas en millones de euros, para el año 2015.

a) Plantee un modelo econométrico que permita obtener la variación del gasto en I+D en millones de euros ante cambios porcentuales de las ventas.
b) Utilizando la base de datos descrita, estime ahora el modelo planteado. Presente las ecuaciones estimadas de la forma habitual e interprete sus resultados.
c) ¿Cómo cambiarían los resultados del punto *b*) si el gasto en I+D se expresase en euros, en lugar de en millones de euros?

2. El modelo de regresión múltiple

Considere ahora un modelo de regresión lineal múltiple (RLM) con k regresores:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k + u$$

donde el cumplimiento del supuesto de media condicionada nula $E\,(u \mid x_1, x_2, \ldots, x_k) = E\,(u) = 0$ garantiza la validez de las estimaciones de los parámetros de pendiente. En cualquier caso, a diferencia del modelo lineal simple, el parámetro $\beta_j\,(j = 1, 2, \ldots, k)$ del modelo RLM recoge el efecto parcial de x_j sobre y, manteniendo contantes el resto de regresores distintos a x_j que se consideran en el modelo (*ceteris paribus*):

$$\beta_1 = \frac{\Delta y}{\Delta x_1}\bigg|_{\substack{\Delta x_2=0,\ldots,\ \Delta x_k=0 \\ \Delta u=0 \leftarrow \text{Si } E(u|x_1,\, x_1,\ldots,x_k)=E(u)=0}}$$

De nuevo, a partir de una muestra de datos para y_i y los x_{ij}, podemos obtener los valores estimados de los parámetros $\beta_0, \beta_1, \ldots, \beta_k$ utilizando los métodos MM o MCO.

Advertencia: Incorporar más regresores en el modelo ↓ SCE y ↑ R^2. Por este motivo, el R^2 no debe utilizarse para decidir si hay que añadir uno o más regresores al modelo. El criterio que debe seguirse para ello es la inferencia estadística (sección 3).

Supuestos Gauss-Markov del modelo RLM con datos de corte transversal:

- RLM1. Linealidad en los parámetros β (los β son únicamente elevados a 1).
- RLM2. Muestra aleatoria de observaciones para la y y las x.
- RLM3. Media condicionada nula, $E(u \mid x_1, x_2, \ldots, x_k) = E(u) = 0$. Independencia entre las x y todos los demás factores inobservables que pueden explicar y (contenidos en u).
- RLM4. No colinealidad perfecta: (a) ninguna variable explicativa es constante para i, y (b) las variables explicativas no están perfectamente correlacionadas.

- RLM5. Homoscedasticidad: $var\ (u \mid x_1, x_2, \ldots, x_k) = \sigma^2 \rightarrow var\ (u \mid x_1, x_2, \ldots, x_k) = \sigma^2$

El cumplimiento de los RLM1-4 garantiza la propiedad estadística de insesgadez del estimador MCO ($E[\hat{\beta}_j] = \beta_j$), mientras que el RLM5 se añade para garantizar la propiedad estadística de eficiencia (relativa) del estimador MCO (min $var\ (\hat{\beta}_j)$ entre todos los estimadores lineales insesgados).

La omisión de una variable explicativa relevante en el modelo puede comprometer el RLM3 y provocar que el estimador MCO esté sesgado. La inclusión de una variable irrelevante en el modelo no provoca sesgo, pero puede reducir la eficiencia (relativa).

La eficiencia (relativa) o precisión del estimador MCO disminuirá cuando:

$$\uparrow var(\hat{\beta}_j) = \frac{\sigma^2}{SCE_j(1-R_j^2)}$$

- ↑ la varianza del error, $E\left(u - E(u)\right)^2 = \sigma^2$
- ↓ la varianza muestral de x_{ij}, $SCE_j = \sum_{i=1}^{n}\left(x_{ij} - \bar{x}_j\right)^2$
- ↑ el grado de colinealidad entre x_j y las otras x's, R_j^2.

Práctica 2A

Ejercicio 2A.1

De los siguientes modelos, indique cuál/es cumple/n la hipótesis de linealidad en los parámetros y, por tanto, podría/n estimarse por el método de mínimos cuadrados ordinarios (MCO):

a) $y = \beta_0 + \beta_1 x + u$

b) $log(y) = \beta_0 + \beta_1\, x + u$

c) $y = \beta_0 + \sqrt{\beta_1}\, x + u$

d) $y = e^{\beta_0} x^{\beta_1} e^{u}$

e) $y = \beta_1 + \beta_2 x + \beta_3 x^2 + u$

f) $y = \beta_1 + \beta_2 \left(\frac{1}{x}\right) + u$

Ejercicio 2A.2

Considere el modelo de regresión lineal múltiple que presentamos a continuación:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + u$$

Asumiendo que se cumple el supuesto de media condicionada nula:

a) Indique cuál es la variación esperada en y si x_1 aumenta en 5 unidades y x_2 se mantiene constante.
b) Indique cuál es la variación esperada en y si x_2 disminuye 3 unidades y x_1 se mantiene constante.
c) Indique cuál es la variación esperada en y si x_1 aumenta en 5 unidades y x_2 disminuye en 3 unidades.

Ejercicio 2A.3

La base de datos «Data_Palma_Mallorca_alquileres.gdt» contiene información extraída Nestoria (https://www.nestoria.es/) el 27 de agosto del 2018 sobre una muestra de pisos en alquiler situadas en Palma de Mallorca. Utilice la base de datos para estimar el modelo econométrico que se plantea a continuación:

$$precio = \beta_0 + \beta_1\, m2 + \beta_2\, dormitorios + \beta_3\, dist_centro + u$$

donde *precio* es el precio de alquiler expresado en euros mensuales, *m2* la superficie útil del piso expresada en metros cuadrados, *dormitorios*es el número de dormitorios, *dist_centro* es la distancia con respecto al centro de la ciudad (expresada en kilómetros), y *u* representa el término de error.

a) Presente la ecuación estimada de la forma habitual e indique qué porcentaje de la variación en el precio de alquiler viene explicado por la superficie, el número de dormitorios y la distancia con respecto el centro.
b) Indique cuál sería la variación estimada en el precio de alquiler del piso con un dormitorio adicional, manteniendo fija la superficie del piso y la distancia con respecto el centro. Interprete el resultado, ¿tiene sentido?
c) Indique cuál será la variación estimada en el precio de alquiler del piso con un dormitorio adicional de una superficie aproximada de 10 metros cuadrados, que mantiene fija la distancia con respecto el centro. Compare el resultado con su respuesta en el apartado anterior.
d) A partir de la función de regresión muestral, obtenga el precio de alquiler predicho para una vivienda con 110 m^2 de superficie útil, con 2 dormitorios, y situado a 2 km del centro.
e) Suponga que el precio de alquiler de la vivienda descrita en el apartado anterior ha acabado siendo en realidad de 1.500 euros mensuales. Calcule el residuo para esta vivienda. Asumiendo que el modelo estimado es cierto, ¿cree que el alquiler del piso es excesivo?
f) Estime el siguiente modelo econométrico e interprete todos los parámetros estimados:

$$log(precio) = \beta_0 + \beta_1 \log(\text{m2}) + \beta_2 log(dormitorios) + \beta_3 dist_centro + u$$

g) Utilizando la función de regresión muestral basada en el modelo del apartado *f*, obtenga el valor predicho de precio cuando *m2* = 110, *dormitorios* = 2, y *dist_centro* = 2. ¿Es dicha predicción puntual mejor o peor que la obtenida en el apartado *e*)?, ¿por qué?

Práctica 2B

Ejercicio 2B.1

El siguiente modelo suele utilizarse para explicar el salario de los individuos:

$$\text{wage} = \beta_0 + \beta_1 age + \beta_2 educ + \beta_3 exper + u$$

donde *wage* representa el salario en euros/mes, *age* son los años del individuo, *educ* los años de formación totales, y *exper* los años de experiencia en el mercado laboral.

Considere que se dispone de un conjunto de datos con información sobre salarios, edad y años de educación (formación académica) para una muestra de individuos. Desafortunadamente, no se dispone de información sobre la experiencia en el mercado laboral. Así pues, alternativamente, se ha utilizado una medida de experiencia potencial, definida como exper = age – educ – 3 (los individuos generalmente empiezan a los tres años el colegio). Explique por qué en este caso no podrían estimarse los parámetros del modelo planteado.

Ejercicio 2B.2

Razone la veracidad o falsedad de las siguientes afirmaciones:

a) Para poder estimar un modelo econométrico por el método de mínimos cuadrados ordinarios es preciso que este sea lineal en las variables.
b) La correlación entre las variables *x* e *y* permite predecir el valor promedio de *y* conociendo los valores de *x*.
c) Si la variable explicativa (*x*) es constante, no es posible estimar el efecto que la misma tiene sobre la variable dependiente (*y*).
d) Hay una relación positiva entre la varianza del estimador de MCO y el número de observaciones, por lo que al disminuir el tamaño de la muestra puede mejorar la eficiencia de nuestra estimación.
e) La inclusión de un regresor irrelevante puede provocar sesgo en los estimadores MCO si este está correlacionado con el resto de regresores incluidos en el modelo.

Ejercicio 2B.3

Considere el siguiente modelo que explica el precio de venta (expresado en euros) de una lavadora (*plav*) en el mercado de segunda mano en términos del número de usos que ha tenido (*usage*) y de la edad (*age*) de la lavadora:

$$plav = \beta_0 + \beta_1 usage + \beta_2 age + u$$

Suponiendo que dicho modelo satisfaga los supuestos de Gauss-Markov, explique cuál sería el sesgo probable obtenido a partir de una regresión lineal simple de *plav* sobre *usage*.

Ejercicio 2B.4

Se dispone de los siguientes datos sobre las ventas en Lilliput de cinco marcas distintas de teléfonos móviles:

Marcas	*VT*	*PR*	*PB*
Elephone	10	8	5,5
Nikita	8	12	8,5
Saoni	7	13	9,0
Plophon	6	24	12,5
Pepaphone	13	9	6,5

Donde, *VT* son las ventas anuales, expresadas en monedas de oro, *PR* es un índice de precios relativos y *PB* son los gastos anuales en publicidad y campañas de promoción, expresados también en monedas de oro.

Tomando como base la anterior información:

a) Estime por MCO los coeficientes del modelo siguiente: $VT_i = \beta_0 + \beta_1 PR_i + u_i$.
b) Obtenga el coeficiente de determinación de esta regresión e interpreta el valor calculado.
c) Obtenga el coeficiente de correlación entre *PR* y *PB*. ¿Cree que sería conveniente añadir la publicidad como variable explicativa adicional en nuestra regresión para mejorar el ajuste de la misma? Razone su respuesta.

Ejercicio 2B.5

Considere el siguiente modelo que describe el precio €/litro del combustible diésel (*p_goa*) fijado por las gasolineras de una ciudad en función del número de

ISBN: 978-84-17900-23-6
DOI: http://dx.doi.org/10.6035/Sapientia163

rivales cercanos[6] al que se enfrentan (gasolineras de distinta marca o *rivals*), y la distancia en kilómetros con respecto a la refinería-almacén más cercano (*distref*):

$$p_goa = \beta_0 + \beta_1 rivals + \beta_2 \log(distref) + u$$

a) Explique cuáles son los signos esperados de β_1 y β_2.

b) Utilizando una muestra de datos de corte transversal para 597 gasolineras valencianas (i = 1, 2, … , 597) descargada a fecha del 15 de enero del 2017 del Ministerio de Energía, Turismo y Agenda Digital (http://geoportalgasolineras.es/), se ha obtenido la siguiente tabla de resultados:

Modelo 1: MCO, usando las observaciones 1-597
Variable dependiente: *p_goa*

	Coeficiente	Desv. Típica	Estadístico *t*	valor *p*	
const	1,07469	0,0253010	42,48	<0,0001	***
rivals	−0,00578616	0,00150087	−3,855	0,0001	***
l_distref	0,0126157	0,00575150	2,193	0,0287	**

Media de la vble. dep.	1,124983	D.T. de la vble. dep.	0,043738
Suma de cuad. residuos	1,096875	D.T. de la regresión	0,042972
R-cuadrado	0,037939	R-cuadrado corregido	0,034700
F(2, 594)	11,71217	Valor p (de F)	0,000010
Log-verosimilitud	1033,280	Criterio de Akaike	−2060,560
Criterio de Schwarz	−2047,384	Crit. de Hannan-Quinn	−2055,430

Presente la ecuación estimada de la forma habitual e interprete los parámetros estimados por MCO asociados a la constante y a las variables explicativas *rivals* y log(*distref*).

c) Indique qué proporción de la variación total de *p_goa* viene explicada por *rivals* y log(distref). Justifique su respuesta.

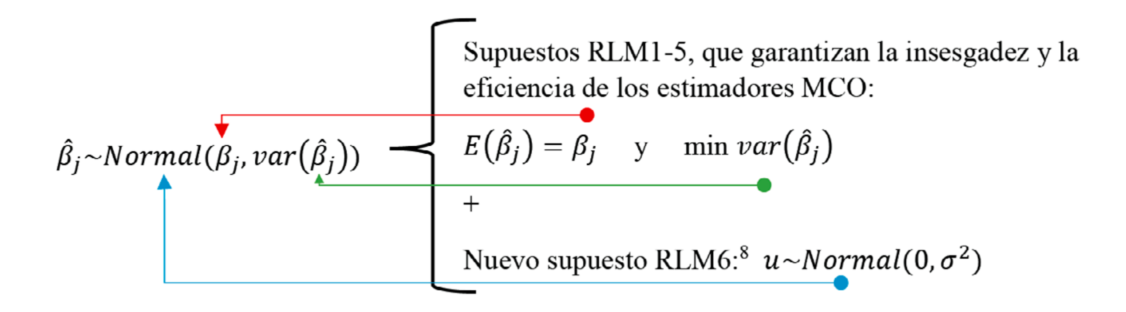
d) Si mantenemos log (*distref*) fija, ¿cuánto tendría que aumentar el número de rivales (*rivals*) para disminuir el precio del combustible en 0,05 €/litro?

6. La cercanía se ha definido trazando un radio de 500 metros sobre cada estación de servicio.

e) Asumiendo que el modelo econométrico inicial cumple con los supuestos de Gauss-Markov y sabiendo que las zonas más alejadas de la refinería-almacén se corresponden mayoritariamente con aquellas zonas con menor densidad de gasolineras, ¿cuál sería el sesgo probable que obtendríamos a partir de una regresión lineal simple de *p_goa* sobre *rivals*?, ¿por qué?

3. Inferencia estadística en modelos de regresión

Deseamos contrastar, a partir de la FRM, hipótesis sobre la población. Para ello, además de conocer var($\hat{\beta}_j$) y E[$\hat{\beta}_j$] de los estimadores MCO, es necesario conocer su distribución muestral,[7] la cual depende de la distribución de u_i.



Podemos llevar a cabo contrastes de hipótesis simples a través del estimador MCO estandarizado (también conocido como estadístico t):

$$\frac{\hat{\beta}_j - \beta_j}{se(\hat{\beta}_j)} \sim t_{n-k-1} \qquad donde \qquad se(\hat{\beta}_j) = \sqrt{\frac{\hat{\sigma}^2}{SCE_j(1 - R_j^2)}} \qquad \frac{\sum_{i=1}^{n} \hat{u}_i^2}{n-k-1}$$

estimación insesgada de $var(u_i|x_i)$

Se distribuye como una t de Student, que depende del tamaño de la muestra, n, y el número de parámetros del modelo distintos de la constante, k.

Su magnitud indica cuántas s.e. difiere la estimación puntual $\hat{\beta}_j$ del valor hipotético $\beta_j = a_j$. El error muestral se considera a través del s.e.

7. La distribución muestral del estimador MCO es la distribución de frecuencias de los valores β_j obtenidos a partir de la estimación de un modelo para cada una de las muestras aleatorias posibles de una población. Conocer dicha distribución muestral nos permitirá obtener, a partir de una sola muestra, la probabilidad de que nuestra estimación se aproxime al parámetro poblacional.
8. Justificación: u aglutina muchos factores inobservables y diferentes. Según el teorema central del límite, la distribución de la suma de un conjunto de variables aleatorias independientes e igualmente distribuidas tiende a ser gaussiana a medida que n crece.

Contrastes de hipótesis simples

1. Planteamos la hipótesis nula y alternativa sobre los parámetros poblacionales:

Contraste de dos colas:
H_0: $\beta_j = a_j$ El efecto de x_j sobre y es igual a a_j, controlado el efecto de las otras x.
H_1: $\beta_j \neq a_j$ El efecto de x_j sobre y no es igual a a_j, controlado el efecto de las otras x.

Contraste de cola derecha:
H_0: $\beta_j = a_j$ El efecto de x_j sobre y es igual a a_j, controlado el efecto de las otras x.
H_1: $\beta_j > a_j$ El efecto de x_j sobre y es mayor que a_j, controlado el efecto de las otras x.

Contraste de cola izquierda:
H_0: $\beta_j = a_j$ El efecto de x_j sobre y es igual a a_j, controlado el efecto de las otras x.
H_1: $\beta_j < a_j$ El efecto de x_j sobre y es menor que a_j, controlado el efecto de las otras x.

2. Construimos el estadístico t para βj: $\hat{\beta}_j$: $t_{\hat{\beta}_j} = \frac{\hat{\beta}_j - a_j}{se(\hat{\beta}_j)} \sim t_{n-k-1}$

Cuando $\hat{\beta}_j$ difiere lo «suficiente» del valor hipotético a_j, considerando el error muestral, se($\hat{\beta}_j$), entonces rechazaremos la H_0. ¿Qué entendemos por suficiente?

3. Elegimos un nivel de significatividad (α), la probabilidad de cometer un error tipo I (rechazar la H_0 cuando realmente es cierta) que estamos dispuestos a asumir en el contraste. Generalmente, α = 0,1, 0,05 o 0,01.
4. $\hat{\beta}_j$ difiere lo «suficiente» de a_j, considerando se($\hat{\beta}_j$), cuando $t_{\hat{\beta}j}$ es más extremo que el valor crítico (c) que define el percentil (1 – α) en la distribución t con n-k-1 grados de libertad.
5. Regla de decisión.

Contraste de dos colas: Rechazamos H_0 cuando $\left|t_{\hat{\beta}_j}\right| > c_{\alpha/2}^{t_{n-k-1}}$

Contraste de cola derecha: Rechazamos H_0 cuando $t_{\hat{\beta}_j} > c_{\alpha}^{t_{n-k-1}}$

Contraste de cola izquierda: Rechazamos H_0 cuando $t_{\hat{\beta}_j} < c_{\alpha}^{t_{n-k-1}}$

6. Concluimos, indicando a qué nivel α se ha realizado el contraste.
En vez de elegir un nivel de significatividad, alternativamente puede utilizarse el p-valor. Considerando el estadístico $t_{\hat{\beta}j}$ obtenido, el p-valor es el nivel de significatividad más pequeño al que se rechazaría la H_0. Cuando p-valor ≤α → rechazamos H_0 a un nivel de significatividad α.

ISBN: 978-84-17900-23-6
DOI: http://dx.doi.org/10.6035/Sapientia163

Intervalos de confianza

Bajo los supuestos del RLM 1 – 6, podemos construir un intervalo de confianza (IC) para el parámetro poblacional β_j:

$$Pr\left(\hat{\beta}_j - c_{\frac{\alpha}{2}}^{t_{n-k-1}} se(\hat{\beta}_j) \leq \beta_j \leq \hat{\beta}_j + c_{\frac{\alpha}{2}}^{t_{n-k-1}} se(\hat{\beta}_j)\right) = (1-\alpha)$$

El límite inferior, $\hat{\beta}_j - c_{\frac{\alpha}{2}}^{t_{n-k-1}} se(\hat{\beta}_j)$, y el límite superior, $\hat{\beta}_j + c_{\frac{\alpha}{2}}^{t_{n-k-1}} se(\hat{\beta}_j)$, albergan el valor poblacional β_j en el 100 × (1-α) % de todas las muestras aleatorias posibles. Así pues, el IC alberga todos los valores para los que la H_0: $\beta_j = a_j$ no podría rechazarse a un α (versus H_1:$\beta_j \neq a_j$) . Útil para realizar contrastes *bilaterales*.

Contraste de una combinación lineal de parámetros

Sobre el modelo y = $\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + u$, se desea realizar el siguiente contraste:

H_0: β_1 - $\beta_2 = a_j$ La diferencia de efectos de x_1 y x_2 sobre y es igual a a_j, controlado el efecto de x_3.

H_0: β_1 - $\beta_2 \neq a_j$ La diferencia de efectos de x_1 y x_2 sobre y es igual a a_j, controlado el efecto de x_3.

En estos casos, no podemos proceder del mismo modo que en un contraste de hipótesis simple, ya que las salidas de los programas habituales utilizados en los cursos de Iniciación a la Econometría no proporcionan toda la información necesaria para construir el estadístico t para una combinación lineal de parámetros:

$$t_{\hat{\beta}_1+\hat{\beta}_2} = \frac{\hat{\beta}_1-\hat{\beta}_2 - a_j}{se(\hat{\beta}_1-\hat{\beta}_2)} \sim t_{n-k-1} \quad \text{donde} \quad se(\hat{\beta}_1-\hat{\beta}_2) = \sqrt{var(\hat{\beta}_1-\hat{\beta}_2)} = \sqrt{var(\hat{\beta}_1)+var(\hat{\beta}_2) - 2Cov(\hat{\beta}_1,\hat{\beta}_2)}$$

Así pues, se aconseja en este caso redefinir el modelo procediendo de la siguiente forma alternativa:

1. Redefinimos la combinación lineal de parámetros, β_1-$\beta_2 = \delta_1$, y replanteamos el contraste en consecuencia:

   H_0:$\delta_1 = a_j$ La diferencia de efectos de x_1 y x_2 sobre y (δ_1) es igual a a_j, controlado el efecto de x_3.

   H_0:$\delta_1 \neq a_j$ La diferencia de efectos de x_1 y x_2 sobre y (δ_1) no es igual a a_j, controlado el efecto de x_3.

2. Dado que $\beta_1 - \beta_2 = \delta_1$, sustituimos en el modelo uno de los parámetros originales. Por ejemplo, sustituimos $\beta_1 = \delta_1 + \beta_2$ y reordenamos expresión:

$$y = \beta_0 + (\delta_1 + \beta_2)x_1 + \beta_2 x_2 + \beta_3 x_3 + u$$

$$y = \beta_0 + \delta_1 x_1 + \beta_2(x_1 + x_2) + \beta_3 x_3 + u$$

3. Estimamos por MCO el modelo reparametrizado que acabamos de presentar:

$$MCO \rightarrow \quad \underset{se(\hat{\beta}_0)}{\hat{y} = \hat{\beta}_0} + \underset{se(\hat{\delta}_1)}{\hat{\delta}_1 x_1} + \underset{se(\hat{\beta}_2)}{\hat{\beta}_2(x_1 + x_2)} + \underset{se(\hat{\beta}_3)}{\hat{\beta}_3 x_3}$$
$$n \qquad R^2$$

4. Con la FRM del modelo reparametrizado, ahora construimos el estadístico t para δ1, el cual nos permitirá llevar a cabo el contraste de la combinación lineal de parámetros planteada, dado que $\beta_1 - \beta_2 = \delta_1$…

$$t_{\hat{\delta}_1} = \frac{\hat{\delta}_1 - a_j}{se(\hat{\delta}_1)} \sim t_{n-k-1}$$

Contraste de hipótesis múltiples

Considerando el siguiente modelo RLMN $y = \beta_0 + \beta_1 x_1 + + \beta_3 x_3 + u$, ahora imagine que desea contrastar múltiples hipótesis sobre los parámetros:

H_0: $\beta_1 = 0, \beta_2 = 0$. x_1 y x_2 no tienen efecto conjuntamente significativo sobre y, una vez controlado el efecto de x_3.
H_1: H_0 no es cierta. x_1 y x_2 tienen efecto conjuntamente significativo sobre y, una vez controlado el efecto de x_3.
Modelo no restringido (nr): $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + u$
Modelo restringido (r) por la H_0: $y = \beta_0 + \beta_3 x_3 + \varepsilon$
(modelo únicamente verdadero si la H_0 es cierta)
Si la H_0 no es cierta, pasar del modelo nr al modelo r dará lugar a que el ajuste de la regresión empeore: $SCE_{nr} < SCE_r$.[9] Así pues, la inferencia en este caso se basa en la tasa de variación de las SCE de pasar de un modelo nr a otro r, ajustado por sus respectivos g.d.l.:[10]

$$F = \frac{(SCE_r - SCE_{nr})/q}{SCE_{nr}/(n-k-1)} \sim F_{q,\ n-k-1}$$

9. $SCE_{nr} = \sum_i^n (y_i - \hat{y}_i)^2 = \sum_{i=1}^n \hat{u}_i^2$ y $SCE_r = \sum_i^n (y_i - \tilde{y}_i)^2 = \sum_{i=1}^n \tilde{\epsilon}^2$.
10. Diferencia de g.d.l. entre modelos r y nr = (n-(k-q)-1-n+k+1) = q.

ISBN: 978-84-17900-23-6
DOI: http://dx.doi.org/10.6035/Sapientia163

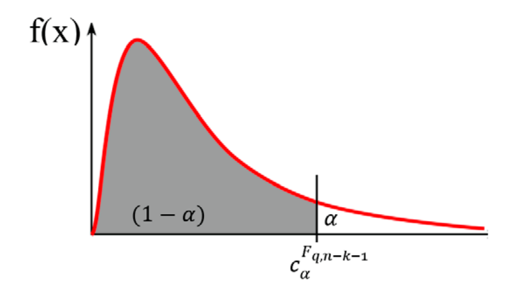
Bajo la H_0, el estadístico F se distribuye como una F de Snedecor con q y n-k-1 g.d.l.[11] ¿Por qué? Bajo el supuesto RLM1 – 6, la SCE es la suma de elementos distribuidos como una normal…

$$F = \frac{\left(\sum_{i=1}^{n} \tilde{\epsilon}^2 - \sum_{i=1}^{n} \widehat{u_i}^2\right)/q}{\sum_{i=1}^{n} \widehat{u_i}^2 / (n-k-1)} \sim \frac{\mathcal{X}_g^2/q}{\dfrac{\mathcal{X}_{n-k-1}^2}{n-k-1}} \sim F_{q,\ n-k-1}$$

donde X_g^2 son distribuciones independientes Chi-cuadrado con g grados de libertad (g.d.l.).

¿Cuándo F es «suficientemente» grande como para rechazar la H_0?

- Elegimos el nivel de significatividad α
- Rechazamos H_0 cuando el estadístico F sea más extremo que el valor crítico (c) que marca el percentil $(1 - \alpha)$ de una distribución F con q y $(n - k - 1)$ g.d.l. en el numerador y denominador, respectivamente.



Téngase en cuenta que, en aquellos casos en los que los modelos sin restringir y restringido tienen la misma variable dependiente, el estadístico F puede expresarse también en términos de los coeficientes de determinación, R^2, de cada uno de los modelos:

$$R^2 = 1 - \frac{SCE}{STC}; \quad SCE = (1 - R^2)STC$$

$$F = \frac{(SCE_r - SCE_{nr})/q}{SCE_{nr}/(n-k-1)} = \frac{(R_{nr}^2 - R_r^2)/q}{(1 - R_{nr}^2)/(n-k-1)}$$

Advertencia: Determinadas restricciones (p. ej., H0: $\beta_1 = 1$, $\beta_2 = 0$) pueden alterar la variable dependiente del modelo restringido, imposibilitando por tanto la utilización de esta última expresión.

11. Este estadístico F puede utilizarse de forma similar, mediante la comparación del modelo no restringido y modelo restringido, para contrastar restricciones lineales como las señaladas en el apartado anterior, resultado de una combinación lineal de los parámetros. Por ejemplo, considerando la hipótesis nula H_0: $\beta_1 - \beta_2 = a_j$, la versión restringida del modelo $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + u$ sería $y = \beta_0 + \text{aj} + \beta_2 x_1 + \beta_2 x_2 + \beta_3 x_3 + \varepsilon \rightarrow y - a_j = \beta_0 + x_1 + \beta_2(x_2 + x_1) + \beta_3 x_3 + \varepsilon$.

ISBN: 978-84-17900-23-6
Jordi Ripollés Piqueras, Inmaculada Martínez Zarzoso
y Maite Alguacil Marí
DOI: http://dx.doi.org/10.6035/Sapientia163

Práctica 3A

Ejercicio 3A.1

Considere el siguiente modelo: $y_i = \beta_0 + \beta_1 x_{i1} + u_i$, donde y_i y x_{i1} representan, respectivamente, la nota de Fundamentos de Econometría y la nota media de la carrera de un conjunto de estudiantes i=1, 2,…, n. Con la muestra de 100 estudiantes del archivo «Data_marks.xlsx» se han obtenido los siguientes resultados:

$$STC = \sum_{i=1}^{n} (y_i - \bar{y})^2 = 467{,}558 \qquad STC_1 = \sum_{i}^{n} (x_{i1} - \bar{x}_1)^2 = 41{,}594$$

$$SEC = \sum_{i=1}^{n} \left(\hat{y}_i - \bar{\hat{y}}\right)^2 = 145{,}345 \qquad \bar{x} = 6{,}417$$

$$SCE = \sum_{i=1}^{n} \hat{u}_i^2 = 322{,}213 \qquad \bar{y} = 5{,}310$$

$$\sum_{i}^{n} (x_{i1} - \bar{x}_1)(y_i - \bar{y}) = \sum_{i}^{n} (x_{i1} - \bar{x}_1) y_i = 77{,}752$$

a) Con la información disponible, obtenga e interprete los parámetros estimados por MCO del modelo planteado.

b) ¿Tiene la nota media de la carrera un efecto sobre la nota de Fundamentos de Econometría? Plantee y realice el contraste a un nivel de significatividad del 5 %. (Nota: necesita calcular los errores estándar, s.e., asociados a los parámetros estimados a partir de la información disponible.)

c) Considere un modelo distinto, donde se han incluido como regresores adicionales las horas que cada estudiante ha invertido en estudiar el examen de la asignatura (x_{i2}) y el número de convocatorias que han consumido (x_{i3}). Utilizando la muestra de 100 estudiantes, se conoce la siguiente información:

$$\hat{y}_i = -1{,}597 + 0{,}382 x_{i1} + 0{,}042 x_{i2} + 0{,}142 x_{i3}$$
$$(0{,}698) \quad (\quad) \quad (\quad) \quad (\quad)$$
$$n = 100 \qquad R^2 = 0.9206$$

$$STC_1 = \sum_{i}^{n} (x_{i1} - \bar{x}_1)^2 = 41{,}594 \qquad R_1^2 = 0{,}2598$$

$$STC_2 = \sum_{i}^{n} (x_{i2} - \bar{x}_2)^2 = 211.091{,}508 \qquad R_2^2 = 0{,}2486$$

$$STC_3 = \sum_{i}^{n} (x_{i3} - \bar{x}_3)^2 = 21{,}310 \qquad R_3^2 = 0{,}0269$$

$$SCE = \sum_{i=1}^{n} \hat{u}_i^2 = 37{,}117$$

Utilizando los datos disponibles, obtenga la información faltante referente a los s.e. y, posteriormente, plantee y realice un contraste de significatividad individual sobre el efecto parcial de las horas de estudio sobre la nota de la asignatura.

Ejercicio 3A.2

El siguiente modelo permite evaluar si, durante el periodo analizado (1980-2016), un conjunto de países han convergido en el tiempo los unos con los otros en términos de renta per cápita o si, por el contrario, han divergido y las diferencias de renta per cápita entre ellos se han agravado.

$$meangr_{i.} = \beta_0 + \beta_1 \ln(GDP_capi_{1980}) + u_i$$

donde

- $meangr_{i.}$ es el promedio a lo largo del tiempo de las tasas de crecimiento anuales del PIB per cápita (en dólares) en cada país i desde 1980 hasta 2016, definida como $\frac{1}{T-1}\sum_{t=1980}^{T}(\ln(GDP_cap_{it}) - \ln(GDP_cap_{it-1}))$
- $\ln(GDP_cap_{i1980})$ es el logaritmo del nivel inicial del PIB per cápita de cada país i

Cuando los países más pobres crecen a unas tasas mayores que los países más ricos, entonces todos los países tienden paulatinamente al mismo nivel de renta per cápita en el largo plazo. Esta tendencia de convergencia se evidencia, entonces, cuando existe una relación inversa entre la media temporal de las tasas de crecimiento anuales de PIB per cápita y el nivel inicial de este, $\beta_1 < 0$.

La base de datos «Data_covergence.gdt» (fuente: The World Bank) contiene información sobre el PIB per cápita de 141 países desde 1980 hasta 2016, expresado en dólares constantes. Dichos datos se han empleado para obtener el promedio temporal de las tasas de crecimiento anuales de los PIB per cápita de los países entre 1980 y 2016 ($meangr_{i.}$), así como el correspondiente nivel del PIB per cápita inicial en 1980, expresado en logaritmos ($\ln(GDP_cap_{i1980})$). Utilizando dicha base de datos, responda las siguientes preguntas:

a) Compare los estadísticos descriptivos del PIB per cápita de 1980 y 2016.
b) Estime por MCO la ecuación del modelo planteado en el enunciado, presente los resultados de la forma habitual e interprete el parámetro estimado asociado a $\ln(GDP_cap_{i1980})$.
c) Contraste la hipótesis nula $H_0: \beta_1 = 0$ versus la hipótesis alternativa de convergencia $H_1: \beta_1 < 0$. Lleve a cabo el contraste a un nivel de significatividad del 5 %.

ISBN: 978-84-17900-23-6

DOI: http://dx.doi.org/10.6035/Sapientia163

Ejercicio 3A.3

El siguiente modelo trata de explicar el número de hijos que tienen los individuos (*ceb*) en función del salario mensual percibido en dólares (*salary_dol*), la edad (*age*), los años de educación o formación (*edyrs*) y los años como fumador (*smokyrs*):

$$ceb = \beta_0 + \beta_1 log(salary_dol) + \beta_2(age) + \beta_3 log(edyrs) + \beta_3 smokyrs + u$$

a) En términos de los parámetros del modelo, plantee la hipótesis nula de que los años como fumador no influyen sobre el número de hijos, una vez tenido en cuenta la influencia del salario, la edad y los años de educación. Especifique como hipótesis alternativa que los años como fumador disminuyen el número de hijos.

b) Utilizando una muestra de 462 residentes en Colombia en 2009 extraída del *Latin American Migration Project*, se ha obtenido la siguiente función de regresión muestral por MCO:

$$\widehat{ceb} = \underset{(0.982)}{-5.506} + \underset{(0.067)}{0.005}\, log(salary_dol) + \underset{(0.236)}{2.335}\, log(age) - \underset{(0.127)}{0.459}\, log(edyrs) - \underset{(0.004)}{0.005}\, smokyrs$$
$$n = 462 \quad R^2 = 0{,}249$$

¿Cuál es la diferencia estimada en número de hijos entre una persona que nunca ha fumado y otra que lleva 60 años haciéndolo, para valores dados de salario, edad y educación?

c) Lleve a cabo el contraste planteado en el apartado *a*) a un nivel de significatividad del 10 %.

d) En base a los resultados, explique si incluiría la variable *smokyrs* en el modelo definitivo para explicar el número de hijos.

Ejercicio 3A.4

Considere que desea estudiar la relación entre la producción (*Y*) de un conjunto de filiales de su propiedad y los factores productivos empleados, capital (*K*) y trabajo (*L*).

a) A partir de la función de producción Cobb-Douglas $Y = AK^{\beta 1}L^{\beta 2}$, plantee un modelo econométrico que sea lineal en los parámetros y, por tanto, pueda ser estimado mediante el procedimiento de MCO.

b) La base de datos «Data_production.xlsx» contiene información para el número de bienes producidos en el 2018 para un conjunto de 100 filiales, en las que se conoce el número de trabajadores y los bienes de capital (equipos o maquinarias) que cada filial ha utilizado para llevar a cabo la producción anual. Utilizando la información disponible, estime los parámetros del modelo propuesto por MCO.

c) Contraste la hipótesis nula de que existen rendimientos constantes a escala, es decir, H_0: $\beta_1 + \beta_2 = 1$, frente la alternativa de que existen rendimientos decrecientes a escala, H_1: $\beta_1+\beta_2 < 1$. Explique el significado de la conclusión alcanzada.

Práctica 3B

Ejercicio 3B.1

Considere un modelo que explica la nota obtenida en Econometría (*nota_econometria*) en función de la nota media en la universidad (*nota_media*), el número de convocatorias presentadas (*convocatoria*) y el número de tutorías que cada alumno ha asistido (*tutorias*). A continuación, presentamos la función de regresión muestral obtenida por MCO para una muestra de 187 estudiantes (Base de datos: Data_marks2):

$$\widehat{nota_econometria} = \underset{(1.222)}{-0.711} + \underset{(0.180)}{0.988}\, nota_media - \underset{(0.256)}{0.226}\, convocatoria + \underset{(0.151)}{0.481}\, tutorias$$

$$n = 187 \quad R^2 = 0.207$$

a) Calcule un intervalo de confianza al 95 % para el parámetro asociado al número de tutorías.

b) Utilizando el intervalo de confianza estimado, contraste si el número de tutorías a las que se asiste influye sobre la nota obtenida en la asignatura de Econometría a un nivel de confianza del 95 %.

c) Contraste si $\beta_{tutorías} = 0{,}5$ a un nivel de confianza del 95 %.

Ejercicio 3B.2

Considere la siguiente ecuación que relaciona el precio de alquiler de las viviendas con el número de habitaciones (*dormitorios*), número de baños (*banos*), tamaño de la vivienda en metros cuadrados (*m2*), la distancia con respecto el centro de la ciudad (*dist_centro*), y el número de viviendas Airbnb que se encuentran alrededor (*n_airbnb*):[12]

$$\log(\text{precio}) = \beta_0 + \beta_1 dormitorios + \beta_2 banos + \beta_3 m2 + \beta_4 dist_centro + \beta_5 n_airbnb + u$$

Utilizando la base de datos «Data_Palma_Mallorca_alquileres.gdt», la cual contiene información extraída de Nestoria (https://www.nestoria.com/) e Inside Airbnb (http://insideairbnb.com/), para una muestra de viviendas de alquiler en

12. Se ha considerado un radio de 500 metros alrededor de cada vivienda para definir la cercanía.

Palma de Mallorca en fecha de 27 de agosto del 2019, se ha obtenido la siguiente función de regresión muestral:

$$\widehat{\log(precio)} = \underset{(0{,}078)}{6{,}489} - \underset{(0{,}019)}{0{,}015}\, dormitorios + \underset{(0{,}031)}{0{,}175}\, banos + \underset{(0{,}0004)}{0{,}003}\, m2 + \underset{(0.023)}{0{,}030}\, dist_centro + \underset{(0.0006)}{0{,}001}\, n_airbnb$$

$$n = 348 \qquad SCE = 25{,}763 \qquad R^2 = 0{,}424$$

a) Plantee y realice un contraste que permita evaluar si el número de viviendas Airbnb cercanas es el causante del incremento del precio de los alquileres, *ceteris paribus*.

b) Ahora plantee y realice un contraste que permita evaluar si las características internas de la vivienda son relevantes de forma conjunta (*dormitorios*, *banos* y *m*2). Para ello, sabemos que $\sum_i^n \left(\log(precio_i) - \hat{\beta}_0 - \hat{\beta}_4 dist_centro_i - \hat{\beta}_5 n_airbnb_i\right)^2 = 45{,}008$.

Ejercicio 3B.3

Considere el siguiente modelo que describe el precio €/litro del diesel (*p_goa*) fijado por las gasolineras de una ciudad en función del número de rivales cercanos a los que se enfrentan (gasolineras de distinta marca o *rivals*), el número de gasolineras cercanas de la misma marca (*samebrand*) y la distancia en kilómetros con respecto a la refinería/almacén más cercano (*distref*):[13]

$$p_goa = \beta_0 + \beta_1 rivals + \beta_2 samebrand + \beta_3 log(distref) + u$$

Utilizando la base de datos de corte transversal para 597 gasolineras valencianas (i = 1, 2, …, 597), descargada a fecha del 15 de febrero del 2017 del Ministerio de Energía, Turismo y Agenda Digital (http://geoportalgasolineras.es/), se ha obtenido la siguiente tabla de resultados:

Modelo 1: MCO, usando las observaciones 1-597
Variable dependiente: *p_goa*

	Coeficiente	Desv. Típica	Estadístico *t*	valor *p*	
const	1,07476	0,0247327	43,46	<0,0001	***
rivals	-0,00498577	0,00147476	-3,381	0,0008	***
samebrand	0,0319987	0,00598194	5,349	<0,0001	***
log(distref)	0,0117751	0,00562450	2,094	0,0367	**

13. La cercanía se ha definido trazando un radio de 500 metros sobre cada estación de servicio.

ISBN: 978-84-17900-23-6
DOI: http://dx.doi.org/10.6035/Sapientia163

Media de la vble. dep.	1,124983	D.T. de la vble. dep.	0,043738
Suma de cuad. residuos	1,046384	D.T. de la regresión	0,042007
R-cuadrado	0,082224	R-cuadrado corregido	0,077581
F(3, 593)	17,70914	Valor p (de F)	5,08e-11
Log-verosimilitud	1047,347	Criterio de Akaike	-2086,694
Criterio de Schwarz	-2069,126	Crit. de Hannan-Quinn	-2079,853

a) Interprete el coeficiente estimado asociado a log(*distref*).

b) Contraste la hipótesis de que el precio del diésel fijado por las gasolineras no cambia con la distancia con respecto a la refinería/almacén, contra la alternativa de que incrementa. Realice el contraste al 1, 5 y 10 % de significatividad.

c) Un informe de la Comisión Nacional de los Mercados y la Competencia afirma que el precio del diésel aumenta en 0,05 euros/litro por cada gasolinera cercana adicional de la misma marca que se instala. A partir de los resultados del modelo estimado, construya un intervalo de confianza al 95 % para el aumento del precio por cada gasolinera adicional cercana de la misma marca y utilícelo para contrastar la afirmación del informe.

d) Establezca la hipótesis nula de que el efecto sobre el precio de la entrada de una gasolinera cercana rival más es compensado por el efecto de la entrada de una gasolinera cercana de la misma marca. ¿Por qué no puede usar los resultados del enunciado para probar la hipótesis planteada? Especifique un modelo que proporcione directamente el estadístico t que permita probar la hipótesis planteada y explique cómo realizaría el contraste.

e) Plantee y realice un contraste de significación conjunta de la regresión anterior a un nivel de significatividad del 1 %.

4. Otras cuestiones relacionadas con modelos de regresión

Cambio de escala en las variables

A continuación, vamos a evaluar los cambios que sufre una FRM ante cambios de escala[14] en las variables.

A. Modelo nivel-nivel estimado: $\underset{se(\hat{\beta}_0)}{\hat{y}_A = \hat{\beta}_0} + \underset{se(\hat{\beta}_1)}{\hat{\beta}_1 x_{1A}} + \underset{se(\hat{\beta}_2)}{\hat{\beta}_2 x_{2A}}$, $SCE = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$ $R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$		
Cambio de escala	Función de regresión muestral	Suma de cuadrados residual y coeficiente de determinación
$y_A/c = y_B$	$\hat{y}_B = \hat{\beta}_0/c + \hat{\beta}_1/c\, x_{1A} + \hat{\beta}_2/c\, x_{2A}$ $se(\hat{\beta}_0)/c \quad se(\hat{\beta}_1)/c \quad se(\hat{\beta}_2)/c$	$\frac{SCE}{c^2}$ R^2
$y_A \cdot c = y_B$	$\hat{y}_B = \hat{\beta}_0 c + \hat{\beta}_1 c\, x_{1A} + \hat{\beta}_2 c\, x_{2A}$ $se(\hat{\beta}_0)c \quad se(\hat{\beta}_1)c \quad se(\hat{\beta}_2)c$	$SCE \cdot c^2$ R^2
$x_{1A} \cdot c = x_{1B}$	$\hat{y}_A = \hat{\beta}_0 + \hat{\beta}_1/c\, x_{1B} + \hat{\beta}_2 x_{2A}$ $se(\hat{\beta}_0) \quad se(\hat{\beta}_1)/c \quad se(\hat{\beta}_2)$	SCE R^2
$x_{1A}/c = x_{1B}$	$\hat{y}_A = \hat{\beta}_0 + \hat{\beta}_1 c\, x_{1B} + \hat{\beta}_2 x_{2A}$ $se(\hat{\beta}_0)\, se(\hat{\beta}_1)c \quad se(\hat{\beta}_2)$	SCE R^2

14. Nota: utilizamos *c* para denotar una constante distinta de 0.

B. Modelo nivel-log estimado: $\underset{se(\hat{\beta}_0)}{\hat{y}_A = \hat{\beta}_0} + \underset{se(\hat{\beta}_1)}{\hat{\beta}_1 \log(x_{1A})} + \underset{se(\hat{\beta}_2)}{\hat{\beta}_2 \log(x_{2A})}$, $SCE = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$ $R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i-\hat{y}_i)^2}{\sum_{i=1}^{n}(y_i-\bar{y})^2}$		
Cambio de escala	Función de regresión muestral	Suma de cuadrados residual y coeficiente de determinación
$y_A/c = y_B$	$\hat{y}_B = \hat{\beta}_0/c + \hat{\beta}_1/c \log(x_{1A}) + \hat{\beta}_2/c \log(x_{2A})$ $se(\hat{\beta}_0)/c \quad se(\hat{\beta}_1)/c \quad se(\hat{\beta}_2)/c$	$\frac{SCE}{c^2}$ R^2
$y_A \cdot c = y_B$	$\hat{y}_B = \hat{\beta}_0 c + \hat{\beta}_1 c \log(x_{1A}) + \hat{\beta}_2 c \log(x_{2A})$ $se(\hat{\beta}_0)c \quad se(\hat{\beta}_1)c \quad se(\hat{\beta}_2)c$	$SCE \cdot c^2$ R^2
$\log(x_{1A} \cdot c) = \log(x_{1B})$	$\hat{y}_A = [\hat{\beta}_0 + \hat{\beta}_1 \log(c)] + \hat{\beta}_1 \log(x_{1A}) + \hat{\beta}_2 x_{2A}$ $se(\hat{\beta}_0 + \hat{\beta}_1 \log(c)) \quad se(\hat{\beta}_1) \quad se(\hat{\beta}_2)$	SCE R^2
$\log(x_{1A}/c) = \log(x_{1B})$	$\hat{y}_A = [\hat{\beta}_0 - \hat{\beta}_1 \log(c)] + \hat{\beta}_1 \log(x_{1A}) + \hat{\beta}_2 x_{2A}$ $se(\hat{\beta}_0 - \hat{\beta}_1 \log(c)) \quad se(\hat{\beta}_1) \quad se(\hat{\beta}_2)$	SCE R^2

C. Modelo log-nivel estimado: $\underset{se(\hat{\beta}_0)}{\log(\hat{y}_A) = \hat{\beta}_0} + \underset{se(\hat{\beta}_1)}{\hat{\beta}_1 x_{1A}} + \underset{se(\hat{\beta}_2)}{\hat{\beta}_2 x_{2A}}$, $SCE = \sum_{i=1}^{n}(\log(y_i) - \log(\hat{y}_i))^2$ $R^2 = 1 - \frac{\sum_{i=1}^{n}(\log(y_i)-\log(\hat{y}_i))^2}{\sum_{i=1}^{n}(\log(y_i)-\log(y))^2}$		
Cambio de escala	Función de regresión muestral	Suma de cuadrados residual y coeficiente de determinación
$log(y_A/c) = \log(y_B)$	$\log(\hat{y}_B) = [\hat{\beta}_0 - \log(c)] + \hat{\beta}_1 x_{1A} + \hat{\beta}_2 x_{2A}$ $se(\hat{\beta}_0 - \log(c)) \quad se(\hat{\beta}_1) \quad se(\hat{\beta}_2)$	SCE R^2
$log(y_A \cdot c) = \log(y_B)$	$\log(\hat{y}_B) = [\hat{\beta}_0 + \log(c)] + \hat{\beta}_1 x_{1A} + \hat{\beta}_2 x_{2A}$ $se(\hat{\beta}_0 + \log(c)) \quad se(\hat{\beta}_1) \quad se(\hat{\beta}_2)$	SCE R^2
$x_{1A} \cdot c = x_{1B}$	$\log(\hat{y}_A) = \hat{\beta}_0 + \hat{\beta}_1/c \; x_{1B} + \hat{\beta}_2 x_{2A}$ $se(\hat{\beta}_0) \quad se(\hat{\beta}_1)/c \quad se(\hat{\beta}_2)$	SCE R^2
$x_{1A}/c = x_{1B}$	$\log(\hat{y}_A) = \hat{\beta}_0 + \hat{\beta}_1 c \, x_{1B} + \hat{\beta}_2 x_{2A}$ $se(\hat{\beta}_0) \quad se(\hat{\beta}_1)c \quad se(\hat{\beta}_2)$	SCE R^2

D. Modelo log-log estimado: $\underset{se(\hat{\beta}_0)}{\log(\hat{y}_A) = \hat{\beta}_0} + \underset{se(\hat{\beta}_1)}{\hat{\beta}_1 \log(x_{1A})} + \underset{se(\hat{\beta}_2)}{\hat{\beta}_2 \log(x_{2A})}$, $SCE = \sum_{i=1}^{n}(\log(y_i) - \log(\hat{y}_i))^2$ R^2		
Cambio de escala	Función de regresión muestral	Suma de cuadrados residual y coeficiente de determinación
$log(y_A/c) = \log(y_B)$	$\log(\hat{y}_B) = [\hat{\beta}_0 - \log(c)] + \hat{\beta}_1 \log(x_{1A}) + \hat{\beta}_2 \log(x_{2A})$ $se(\hat{\beta}_0 - \log(c)) \quad se(\hat{\beta}_1) \quad se(\hat{\beta}_2)$	SCE R^2
$log(y_A \cdot c) = \log(y_B)$	$\log(\hat{y}_B) = [\hat{\beta}_0 + \log(c)] + \hat{\beta}_1 \log(x_{1A}) + \hat{\beta}_2 \log(x_{2A})$ $se(\hat{\beta}_0 + \log(c)) \quad se(\hat{\beta}_1) \quad se(\hat{\beta}_2)$	SCE R^2
$\log(x_{1A} \cdot c) = \log(x_{1B})$	$\log(\hat{y}_A) = [\hat{\beta}_0 + \hat{\beta}_1 \log(c)] + \hat{\beta}_1 \log(x_{1A}) + \hat{\beta}_2 \log(x_{2A})$ $se(\hat{\beta}_0 + \hat{\beta}_1 \log(c)) \quad se(\hat{\beta}_1) \quad se(\hat{\beta}_2)$	SCE R^2
$\log(x_{1A}/c) = \log(x_{1B})$	$\log(\hat{y}_A) = \hat{\beta}_0 - \hat{\beta}_1 \log(c) + \hat{\beta}_1 \log(x_{1A}) + \hat{\beta}_2 \log(x_{2A})$ $se(\hat{\beta}_0 - \hat{\beta}_1 \log(c)) \quad se(\hat{\beta}_1) \quad se(\hat{\beta}_2)$	SCE R^2

Las interacciones entre variables explicativas en el modelo de regresión nos permiten modular los efectos marginales. Por ejemplo, cuando interaccionamos dos variables explicativas diferentes, permitimos que el efecto marginal de una variable (x_1) sobre la variable dependiente (y) dependa de otra variable (x_2):

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + u, \qquad \frac{\Delta y}{\Delta x_1} = \beta_1 + \beta_3 x_2$$

Bondad de ajuste y selección de modelos

Anteriormente ya mencionamos que incorporar más regresores en el modelo, sean relevantes o no, ↓ *SCE* y ↑ R^2. Por este motivo, el R^2 no debe utilizarse para decidir si hay que agregar uno o más regresores en el modelo.

Solución:

- La inferencia estadística (contrate t o F, según el caso) es particularmente conveniente para seleccionar modelos con o sin constante y, en general, para comparar modelos anidados. Véase la Sección 3. Inferencia estadística en modelos de regresión.
- Comparar el coeficiente de determinación ajustado ($\bar{R}^2$) es conveniente para elegir entre distintos modelos no anidados, *siempre y cuando se tenga la misma variable dependiente* (y) *y el mismo tamaño muestral*(n).

$$\bar{R}^2 = 1 - \frac{\frac{SCE}{n-k-1}}{\frac{STC}{n-1}}$$

El R^2 es especialmente útil cuando se quiere decidir entre regresores alternativos que recogen un determinado aspecto que puede explicar la variable dependiente, o entre regresores que representan formas funcionales diferentes.

Práctica 4

Ejercicio 4.1

Considere el siguiente modelo que describe la relación entre el nivel de contaminación de los países y su renta (relación Kuznet medioambiental), donde CO2c representa el nivel de contaminación medido en emisiones de CO_2 en toneladas métricas per cápita, y GDPc es el producto interior bruto (PIB) expresado en millones de dólares PPP per cápita.

$$CO2c = \beta_0 + \beta_1 GDPc + \beta_2 GDPc^2 + u$$

A partir de la información disponible en la web https://data.worldbank.org del Banco Mundial, se han descargado datos de corte transversal sobre las variables de interés correspondientes al año 2014 para una muestra de 182 países (i = 1, 2, …, 182). A continuación presentamos una tabla con los principales resultados de la estimación:

Modelo 1: MCO, usando las observaciones 1-182
Variable dependiente: CO2c

	Coeficiente	Desv. Típica	Estadístico *t*	Valor *p*	
const	−0,417032	0,341334	−1,2218	0,2234	
GDPc	351,913	36,9375	9,5273	<0,0001	***
GDPc_sq	−3161,38	696,8	−4,5370	<0,0001	***

Media de la vble. dep.	3,738123	D.T. de la vble. dep.	3,792689
Suma de cuad. residuos	968,7391	D.T. de la regresión	2,326360
R-cuadrado	0,627922	R-cuadrado corregido	0,623765
F(2, 179)	151,0410	Valor p (de F)	3,73e-39
Log-verosimilitud	−410,3978	Criterio de Akaike	826,7956
Criterio de Schwarz	836,4076	Crit. de Hannan-Quinn	830,6921

a) Calcule e interprete el efecto marginal de la renta sobre el nivel de contaminación. Además, según el modelo estimado, represente gráficamente la relación entre la contaminación y la renta, cuantificando el punto de origen de la FRM, su pendiente y posible punto de inflexión.

b) Según los resultados estimados, ¿cree que el modelo debería contener el término cuadrático como regresor? Justifique su respuesta.

c) El PIB de Noruega y España en 2014 fue, respectivamente, 0,066 y 0,034 millones de dólares per cápita. En base a los resultados obtenidos, indique si en estos dos países las políticas de desarrollo económico podrían perjudicar al medioambiente. Razone su respuesta.

d) Sabiendo que 1 dólar = 0,89 euros, escriba la ecuación estimada (incluyendo los errores estándar y el R-cuadrado) que resultaría si expresásemos el PIB en millones de euros per cápita, en vez de millones de dólares per cápita.

e) A continuación presentamos tres ecuaciones estimadas a partir de la base de datos descrita anteriormente. Indique cuál de los modelos planteados sería preferible. Razone su respuesta.

$\widehat{CO2c} = -0{,}417 + 351{,}913\ \text{GDPc} - 3161{,}38\ \text{GDPc}^2$
(0.341) (36,938) (696,8) $R^2 = 0{,}628$ $\bar{R}^2 = 0{,}624$ $n = 182$

$\widehat{CO2c} = 0{,}620 + 192{,}628\ \text{GDPc}$
(0,267) (192,628) $R^2 = 0{,}585$ $\bar{R}^2 = 0{,}583$ $n = 182$

$\widehat{CO2c} = 15{,}818 + 2{,}610\ \log(\text{GDPc})$
(0,799) (0,168) $R^2 = 0{,}573$ $\bar{R}^2 = 0{,}570$ $n = 182$

Ejercicio 4.2

Busque en la red una muestra de datos de *corte transversal* para dos variables de índole económico-empresarial que crea que puedan presentar una relación cuadrática.[15] Utilizando dicha muestra de datos, realice las siguientes tareas:

a) Utilizando Excel, guarde en columnas las dos variables, nombrándolas y ordenándolas, junto a una variable índice i = 1, 2, ... , N para representar la dimensión de corte transversal (p. ej., personas, países, hogares, empresas, etc.). No olvide indicar la fuente de donde se han obtenidos los datos, así como el significado de cada variable y sus unidades de medida.
b) Utilice la teoría económica o el razonamiento lógico para justificar qué variable es la dependiente y cuál/es la/s explicativa/s.
c) Utilizando Excel o Gretl, represente e interprete un diagrama de dispersión (gráfico X-Y) que muestre la relación entre las variables (véase la solución del ejercicio 0.1).
d) Utilizando Gretl, estime por MCO el siguiente modelo: $y = \beta_0 + \beta_1 x_1 + \beta_2 x_1^2 + u$ y exprese la función de regresión muestral resultante de la forma habitual. De acuerdo con los resultados obtenidos, explique cuál es el efecto marginal estimado de x_1 y. Calcule e interprete el punto de inflexión en la relación entre y y x_1.
e) ¿Cree que el modelo debería contener el término cuadrático como regresor? Justifique su respuesta.
f) Ahora someta al regresor x a un cambio de escala de su elección (p. ej., pase de años a meses, de euros a cientos de euros, etc.). Exprese la función de regresión muestral, tras el cambio de escala, explicando cómo cambiarán los coeficientes estimados, los errores estándar, los estadísticos t, el coeficiente de determinación y la SCE.

15. Posibles fuentes de datos: Gapminder (www.gapminder.org), Goolzoom (www.goolzoom.es), Instituto Nacional de Estadística (www.ine.es), Eurostat (https://ec.europa.eu/eurostat/data/database), OECD (https://stats.oecd.org/), World Bank (https://data.worldbank.org/), UNCTADSTAT (https://unctadstat.unctad.org/), FAOSTAT (http://www.fao.org/faostat/en/#data), otros (https://www.economicsnetwork.ac.uk/links/sources, https://db.nomics.world/).

Ejercicio 4.3

Considere las siguientes funciones de regresión muestral:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta} x_2$$

$$\log(\hat{y}) = \hat{\beta}_0 + \hat{\beta}_1 \log(x_1) + \hat{\beta}_2 \log(x_2)$$

$$\log(\hat{y}) = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2$$

a) Explique qué les sucedería a los estimadores MCO, s.e. y R^2, basados en los modelos considerados, si multiplicamos por 100 todos los valores de la variable dependiente y.

b) Suponga, en vez de *a*), que todos los valores de x_1 son multiplicados por 100. Explique cómo se verían afectados en este caso los estimadores MCO, s.e. y R^2.

Ejercicio 4.4

Considere el siguiente modelo econométrico:

$$\log(\text{precio}) = \beta_0 + \beta_1 \log(dorm) + \beta_2 \log(m2) + \beta_3 \log(dorm) \cdot \log(m2) + \beta_4 \log(\text{banos}) + u$$

donde precio es el precio de venta de viviendas expresado en euros, *dorm* el número de dormitorios, *m*2 el tamaño de la vivienda expresado en metros cuadrados, y banos es el número de baños. En base a esta información, conteste a las siguientes preguntas:

a) Según este modelo, ¿cuál es el efecto marginal de los dormitorios sobre el precio?

b) Utilizando la base de datos con información sobre 1.300 viviendas situadas en Castellón que se venden a fecha de 11 de febrero del 2019 (fuente: Nestoria) se ha estimado la siguiente función de regresión muestral:

$$\log(\widehat{precio}) = \underset{(0{,}565)}{9{,}344} - \underset{(0{,}390)}{2{,}877}\log(dorm) + \underset{(0{,}133)}{0{,}463}\log(m2) + \underset{(0{,}089)}{0{,}576}\log(dorm)\cdot\log(m2) + \underset{(0{,}042)}{0{,}608}\log(banos)$$

$$n = 1213 \quad R^2 = 0{,}590$$

Interprete el coeficiente estimado asociado a la interacción. Para ello, considere que el piso tiene 150 metros cuadrados. ¿Y si tuviese 300 metros cuadrados?

c) Contraste si el efecto de los dormitorios sobre el precio depende de los metros cuadrados o no.

5. Incorporación de variables ficticias en el análisis de regresión

Podemos incorporar información cualitativa en el análisis de regresión empleando variables ficticias. Denominamos variable ficticia (o *dummy*) a una variable binaria que toma el valor 1 para denotar que la unidad de corte transversal ipertenece a una determinada categoría, y 0 en caso contrario. Por ejemplo, considere el siguiente modelo econométrico que pretende explicar el precio de los pisos en relación a su tamaño y la presencia (o no) de ascensor:

$$p = \beta_0 + \beta_1 m2 + \beta_2 D_{asc} + u$$

donde *p* es el precio del piso, *m2* representa los metros cuadrados del piso y Dasc es una variable ficticia que toma el valor 1 si el edificio tiene ascensor y 0 si no tiene. Ejemplo:

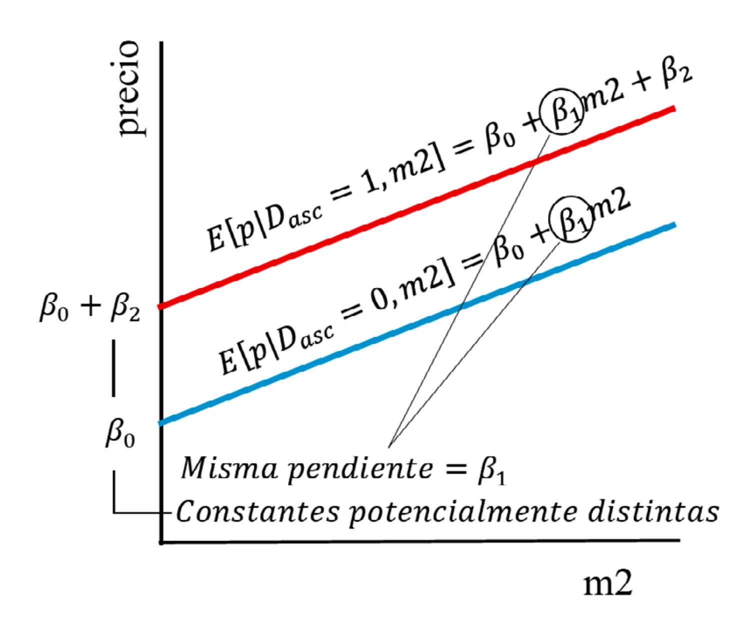


Figura 5.1. Modelo de regresión con una variable ficticia

ISBN: 978-84-17900-23-6
DOI: http://dx.doi.org/10.6035/Sapientia163

β_2 captura la prima del ascensor, es decir, el diferencial de precios medios entre los pisos con versus sin ascensor, para un tamaño dado.

$$\beta_2 = E[p|D_{asc} = 1, \text{m2}] - E[p|D_{asc} = 0, m2]$$

Su relevancia estadística puede contrastarse con el estadístico t para $\beta 2$:

- Si $\beta_2 = 0$ → El ascensor no es importante para determinar el precio del piso, ceteris paribus.
- Si $\beta_2 > 0$ → Los pisos con ascensor son más caros que los pisos sin ascensor, ceteris paribus. Caso representado gráficamente.
- Si $\beta_2 < 0$ → Los pisos con ascensor son más baratos que los pisos sin ascensor, ceteris paribus.

Evite la trampa de la variable dicotómica: si el modelo contiene una constante, la inclusión de una variable dummy por cada posible categoría da lugar a un problema de colinelidad perfecta. Por este motivo, en el ejemplo anterior, se ha incluido únicamente una variable dummy para modelizar dos posibles categorías: con ascensor y sin ascensor.

¿Qué ocurre si tenemos múltiples categorías? Si estuviésemos interesados en evaluar posibles diferencias de precios entre pisos localizados en distintos distritos (p. ej., centro, norte, sur, este y oeste), entonces deberíamos omitir una categoría para evitar un problema de colinealidad perfecta. En ese caso, un modelo válido podría ser el siguiente:

$$p = \beta_0 + \beta_1 m2 + \beta_2 D_c + \beta_3 \text{D}_\text{n} + \beta_4 D_s + \beta_5 D_e + u$$

donde
- D_c toma el valor 1 si el piso está en el distrito centro de la ciudad y 0 en caso contrario
- D_n toma el valor 1 si el piso está en el distrito norte de la ciudad y 0 en caso contrario
- D_s toma el valor 1 si el piso está en el distrito sur de la ciudad y 0 en caso contrario
- D_e toma el valor 1 si el piso está en el distrito este de la ciudad y 0 en caso contrario
- Omitimos la variable dummy de distrito oeste, que representará el grupo de referencia

Así pues, en este caso, los coeficientes asociados a cada variable dummy, β_2, β_3, β_4, β_5, deben interpretarse en términos relativizados respecto a la categoría omitida. Por ejemplo, β_3 nos indicaría la diferencia de precios medios entre los pisos del distrito norte y los del distrito oeste (categoría de referencia omitida):

ISBN: 978-84-17900-23-6

DOI: http://dx.doi.org/10.6035/Sapientia163

$$\beta_3 = E\left[p \middle| \begin{matrix} D_C = 0 \\ D_n = 1 \\ D_s = 0 \\ D_e = 0 \end{matrix}, m2\right] - E\left[p \middle| \begin{matrix} D_C = 0 \\ D_n = 0 \\ D_s = 0 \\ D_e = 0 \end{matrix}, m2\right]$$

El estadístico *t* para $\hat{\beta}_3$ permitiría contrastar si la diferencia de precios medios entre pisos del distrito norte y pisos del distrito oeste es estadísticamente significativa a los niveles habituales.

¿Pero qué ocurre si deseamos conocer la diferencia de precios medios entre pisos del distrito centro y pisos del distrito norte, para un tamaño dado? Ninguna de las dos categorías comparadas es la de referencia (distrito oeste). Aun así, podemos cuantificarlo comparando las funciones de regresión para ambas categorías:

$$E\left[p \middle| \begin{matrix} D_C = 1 \\ D_n = 0 \\ D_s = 0 \\ D_e = 0 \end{matrix}, m2\right] - E\left[p \middle| \begin{matrix} D_C = 0 \\ D_n = 1 \\ D_s = 0 \\ D_e = 0 \end{matrix}, m2\right] = \beta_2(1) - \beta_3(1)$$

Para contrastar si dicha diferencia es estadísticamente significativa, sin embargo, la forma más sencilla sería replantear el modelo y omitir la variable ficticia correspondiente a la categoría respecto de la cual se desea realizar la comparativa:

$$p = \beta_0 + \beta_1 m2 + \delta_2 D_c + \delta_3 D_s + \delta_4 D_e + \delta_5 D_o + u$$

En este caso, el parámetro δ_2 nos indicaría directamente el diferencial de precios del distrito centro y pisos del distrito norte (ahora nuestra categoría de referencia, según el modelo replanteado), para un tamaño dado. Además, el estadístico *t* para δ_2 permitiría realizar el contraste de significatividad correspondiente. Por ejemplo, sen este caso los precios en el distrito centro fuesen menores que en el distrito. Norte ($\delta_2 < 0$), entonces tendríamos lo siguiente:

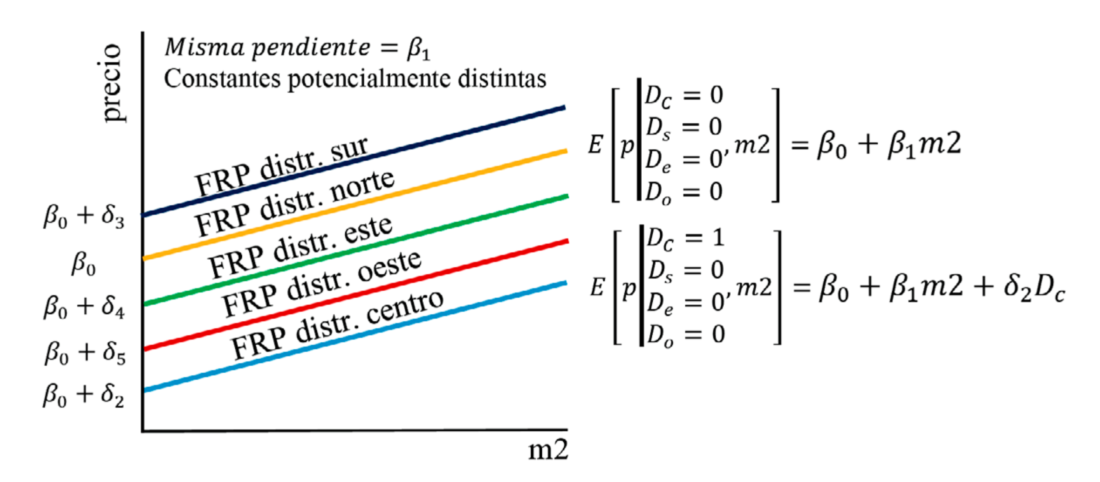


Figura 5.2. Modelo de regresión con múltiples categorías

La interacciónentre una variable continua y una variable ficticia permite que el efecto parcial de la variable continua pueda depender de la pertenencia o no a una determinada categoría. Siguiendo con el primer ejemplo sobre los pisos, ahora interaccionamos la variable continua m2 con la variable ficticia Dasc (= 1 si el piso posee ascensor, 0 en caso contrario):

$$p = \beta_0 + \beta_1 m2 + \beta_2 Dasc + \beta_3 (m2 \cdot Dasc) + u$$

De este modo permitimos que el efecto parcial del tamaño del piso sobre el precio pueda ser distinto dependiendo de si el piso tiene ascensor o no: $\frac{\Delta p}{\Delta m2} = \beta_1 + \beta_3 D_{ac}$.

Si hay ascensor, $\beta_1 + \beta_3$ representa el efecto parcial del tamaño sobre el precio.

Si no hay ascensor, β_1 representa el efecto parcial del tamaño sobre el precio.

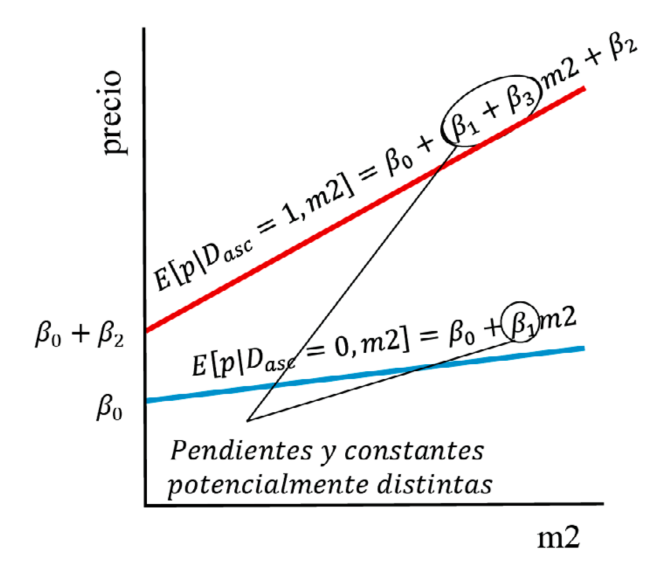


Figura 5.3. Modelo de regresión con una variable ficticia interaccionada con una variable continua

ISBN: 978-84-17900-23-6
DOI: http://dx.doi.org/10.6035/Sapientia163

β_2 captura el diferencial de precios medios entre los pisos con versus sin ascensor, para un tamaño dado.

$$\beta_2 = E[p|D_{asc} = 1, \text{m2}] - \text{E}[\text{p}|\text{D}_{asc} = 0, \text{m2}]$$

Su relevancia estadística es contrastable con el estadístico t para $\hat{\beta}_2$.

β_3 captura la diferencia en el efecto parcial del tamaño sobre el precio entre pisos con y sin ascensor. Su relevancia estadística se contrastaría con el estadístico t para $\hat{\beta}_3$:

- Si $\beta_3 = 0 \rightarrow$ El efecto parcial no depende del ascensor.
- Si $\beta_2 > 0 \rightarrow$ El efecto parcial es mayor cuando hay ascensor. Caso representado gráficamente.
- Si $\beta_2 < 0 \rightarrow$ El efecto parcial es menor cuando hay ascensor.

La interacción entre variables ficticias permite crear categorías diferentes en función de los valores de las ficticias y evaluar las diferencias entre estas categorías. Por ejemplo, considere el siguiente modelo:

$$p = \beta_0 + \beta_1 D_{asc} + \beta_2 D_{terr} + \beta_3 (D_{asc} \cdot D_{terr}) + \gamma_1 m2 + u$$

donde D_{asc} es una variable ficticia que toma el valor 1 si el piso tiene ascensor y 0 en caso contrario, D_{terr} es una variable ficticia que toma el valor 1 si el piso tiene terraza y 0 en caso contrario, y $m2$ representa los metros cuadrados del piso.

Así pues, para un tamaño dado de los pisos, el modelo permitirá obtener el diferencial de precios medios entre las siguientes categorías:

Entre pisos con ascensor y terraza (A), versus pisos sin ascensor ni terraza (B):

$$E[p|D_{asc} = 1, D_{terr} = 1|m2] - E[p|D_{asc} = 0, D_{terr} = 0|m2] = \beta_1 + \beta_2 + \beta_3$$

Entre pisos sin ascensor y con terraza (A), versus pisos sin ascensor ni terraza (B):

$$E[p|D_{asc} = 0, D_{terr} = 1|m2] - E[p|D_{asc} = 0, D_{terr} = 0|m2] = \beta_2$$

Entre pisos con ascensor y sin terraza (A), versus pisos sin ascensor y con terraza (B):

$$E[p|D_{asc} = 1, D_{terr} = 0|m2] - E[p|D_{asc} = 0, D_{terr} = 1|m2] = \beta_1 - \beta_2$$

La relevancia estadística de dichas diferencias es contrastable mediante la correspondiente prueba t, o prueba F, según el caso.

Diferentes FRM según categorías: Finalmente, podemos permitir que, tanto el intercepto como el efecto parcial de todos los regresores (β_j), puedan diferir entre distintas categorías. Para ello, debemos interaccionar todas las variables explicativas

continuas del modelo con una variable ficticia que represente la pertenencia a las categorías de interés. Véase, por ejemplo, la siguiente especificación:

$$p = \beta_0 + \beta_1 m2 + \beta_2\, rooms + \gamma_0 D_{asc} + \gamma_1 (D_{asc} \cdot m2) + \gamma_2 (D_{asc} \cdot rooms) + u$$

donde se ha incluido una variable ficticia Dasc que, además, se encuentra interaccionada con *m*2 y *rooms*. Entonces:

- γ_0 es la diferencia en el intercepto entre pisos con ascensor y pisos sin ascensor,
- γ_1 es la diferencia entre pisos con ascensor y pisos sin ascensor en el efecto marginal del tamaño sobre el precio,
- γ_2 es la diferencia entre pisos con ascensor y pisos sin ascensor en el efecto marginal del número de habitaciones sobre el precio.

Este tipo de especificación es habitualmente utilizada para evaluar cambios estructurales del modelo entre categorías. En nuestro ejemplo, podríamos evaluar si el precio de los pisos sigue el mismo modelo para casos con ascensor y casos sin ascensor, planteando y realizando el siguiente contraste de significatividad conjunta, a través de una prueba F:[16]

H_0: $\gamma_0 = 0$, $\gamma_1 = 0$, $\gamma_2 = 0$ (el precio sigue el mismo modelo en ambas categorías).
H_1: H_0 no es cierta (el precio no sigue el mismo modelo en ambas categorías).

A este contraste se le denomina test de Chow. En Gretl, una vez estimado un modelo por MCO (*Modelo / Mínimos Cuadrados Ordinarios*), puede llevarse a cabo el contraste siguiendo la ruta *Contrastes / Test de Chow* la venta de resultados.

Práctica 5

Ejercicio 5.1

Disponemos de una base de datos que contiene información sobre los salarios medios percibidos por comunidad autónoma (variable *salario*, expresada en euros anuales) y por género (*mujer* una variable ficticia que toma el valor 1 para las mujeres, y 0 para los hombres). Además, la base de datos también contiene dos variables ficticias adicionales: *sur* toma el valor 1 para las comunidades del sur de España, cero en caso contrario; e *islas* toma el valor 1 para las islas, cero en caso contrario. Los datos son para el año 2016 sobre una muestra de trabajadores que residen y trabajan en España. Dicha información ha sido extraída de la Encuesta de Estructura Salarial del Instituto Nacional de Estadística (www.ine.es).

16. Alternativamente, el contraste también podría realizarse estimando el mismo modelo con submuestras.

a) Utilizando la base de datos descrita en el párrafo anterior, interprete los resultados que se han obtenido tras estimar por MCO la siguiente función de regresión muestral:

$$\log(\widehat{salario}) = \underset{(0{,}024)}{10{,}128} - \underset{(0{,}034)}{0{,}261}\, mujer$$
$$n = 34 \qquad R^2 = 0{,}653$$

b) Con los mismos datos se ha estimado el modelo cuyos resultados aparecen a continuación. ¿Cuál es la brecha salarial si el modelo se estima con la variable salario en niveles? Interprete también la constante de la regresión.

Modelo: MCO, usando las observaciones 1-34
Variable dependiente: salario

	Coeficiente	**Desv. Típica**	**Estadístico *t***	**valor *p***	
const	25.159,7	563,300	44,66	<0,0001	***
mujer	−5.808,58	796,627	−7,291	<0,0001	***

Media de la vble. dep.	22.255,40	D.T. de la vble. dep.	3.731,119
Suma de cuad. residuos	1,73e+08	D.T. de la regresión	2.322,545
R-cuadrado	0,624261	R-cuadrado corregido	0,612519
F(1, 32)	53,16548	Valor p (de F)	2,75e-08

c) Un nuevo modelo ha sido estimado incluyendo las variables sur e islas. Indique si el salario medio es significativamente menor en el sur que en el resto de regiones en vista de los resultados obtenidos en el siguiente modelo:

Modelo: MCO, usando las observaciones 1-34
Variable dependiente: l_salario

	Coeficiente	**Desv. Típica**	**Estadístico *t***	**valor *p***	
const	10,1679	0,0226799	448,3	<0,0001	***
sur	−0,116602	0,0342218	−3,407	0,0019	***
mujer	−0,260945	0,0284308	−9,178	<0,0001	***
islas	−0,109062	0,0450551	−2,421	0,0218	**

Media de la vble. dep.	9.997126	D.T. de la vble. dep.	0,163866
Suma de cuad. residuos	0.206119	D.T. de la regresión	0,082889
R-cuadrado	0.767392	R-cuadrado corregido	0,744131
F(3, 30)	32.99079	Valor p (de F)	1,25e-09

Ejercicio 5.2

Considere la siguiente ecuación de gravedad para los flujos de comercio bilateral. Las exportaciones del país *i* al país *j* vienen explicadas por el producto interior bruto (*y*) de *i* y *j*, y por variables geográficas y culturales, entre ellas la distancia geográfica (*D*) entre las capitales de *i* y *j* y tres variables ficticias: frontera común (*contig*) toma el valor 1 si los países comparten frontera y 0 en caso contrario, idioma común (*comlang_off*) toma el valor 1 si los países tienen la misma lengua oficial, 0 en caso contrario y relación colonial (*col_to*) toma el valor 1 si los países tienen o han tenido una relación colonial en el pasado. El siguiente modelo ha sido estimado utilizando el fichero de datos extraído de la base de datos de CEPII «gravity» (www.cepii.fr) para una muestra de 17.088 flujos de exportaciones en 2006 (http://www.cepii.fr):

$$lnX_{ij} = \beta_0 + \beta_1 lnY_i + \beta_2 lnY_j + \beta_3 lnD_{ij} + \beta_4 Contig_{ij} + \beta_5 comlang_off_{ij} + 6col_to_{ij} + u_{ij}$$

Se ha obtenido la siguiente tabla de resultados:

Modelo: MCO, usando las observaciones 1-30569 (n = 17088)
Se han quitado las observaciones ausentes o incompletas: 13481
Variable dependiente: lx

	Coeficiente	Desv. Típica	Estadístico *t*	valor *p*	
const	−10,4228	0,251091	−41,51	<0,0001	***
lnYi	1,24696	0,00853113	146,2	<0,0001	***
lnYj	0,926587	0,00825138	112,3	<0,0001	***
Ln D	−1,36871	0,0250859	−54,56	<0,0001	***
contig	1,16246	0,124409	9,344	<0,0001	***
col_to	0,223536	0,193482	1,155	0,2480	
comlang_off	1,18719	0,0543500	21,84	<0,0001	***

Media de la vble. dep.	1,199532	D.T. de la vble. dep.	4,118644
Suma de cuad. residuos	103530,2	D.T. de la regresión	2,461937
R-cuadrado	0,642815	R-cuadrado corregido	0,642690
F(6, 17081)	5123,366	Valor p (de F)	0,000000
Log-verosimilitud	−39638,73	Criterio de Akaike	79291,46
Criterio de Schwarz	79345,68	Crit. de Hannan-Quinn	79309,33

a) Interprete los coeficientes de las tres variables ficticias. De acuerdo con los resultados obtenidos, ¿comercian más los países que comparten frontera, en comparación con aquellos pares de países que no la comparten? Cuantifique el efecto esperado.

b) Se ha reestimado el modelo añadiendo una interacción entre las variables frontera común (*contig*) e idioma común (*comlang_off*), los coeficientes estimados para *contig* y (*contig**comlang_off*) son respectivamente (desviación típica entre paréntesis):

1,513 -0,892
(0.155) (0,236)

Obtenga el efecto parcial sobre las exportaciones que tiene el hecho de compartir frontera en el modelo ampliado.

c) Se ha reestimado el modelo añadiendo una interacción entre una variable continua (*distancia*) y la relación colonial. El coeficiente estimado para la interacción no es estadísticamente significativo. ¿Cómo podría interpretarse el resultado?

Ejercicio 5.3

Con los datos obtenidos del Banco Mundial (World Bank Doing Business: http://www.doingbusiness.org) se ha utilizado una muestra de empresas para Egipto en 2013 de donde se han obtenido datos de ventas anuales y de número de trabajadores fijos empleados. Además, con esta información se han construido variables ficticias indicando si las empresas de la muestra exportan o no, si están participadas por capital extranjero y de si el mánager principal es una mujer.

a) Interprete los resultados obtenidos en el siguiente modelo de regresión lineal, donde la variable dependiente es el logaritmo de la productividad del trabajo (*llabpro*) y las variables explicativas son: la experiencia del manager (*exper*), la edad de la empresa o años de funcionamiento (*age*) y dos variables ficticias

que indican si la empresa exporta (*exporter*), y si está participada por capital extranjero (*foreign*).

Modelo: MCO, usando las observaciones 1-2897 (n = 2408)
Variable dependiente: llabpro

	Coeficiente	Desv. Típica	Estadístico *t*	valor *p*	
const	11,3169	0,0619118	182,8	<0,0001	***
age	−0,0126260	0,00205075	−6,157	<0,0001	***
exporter	0,605089	0,0743340	8,140	<0,0001	***
exper	0,00567645	0,00271938	2,087	0,0370	**
foreign	0,191541	0,102156	1,875	0,0609	*

Media de la vble. dep.	11,31505	D.T. de la vble. dep.	1,418927
Suma de cuad. residuos	4632,163	D.T. de la regresión	1,388401
R-cuadrado	0,044155	R-cuadrado corregido	0,042564
F(4, 2403)	27,75154	Valor p (de F)	1,47e-22
Log-verosimilitud	−4204,494	Criterio de Akaike	8418,987
Criterio de Schwarz	8447,920	Crit. de Hannan-Quinn	8429,511

b) También se han generado variables ficticias a partir de la variable tamaño de la empresa (*size_cat*), que clasifica las empresas en tres categorías (grande = cat_1, media = cat_2, pequeña = cat_3) según el número de trabajadores. Interprete los resultados obtenidos tras estimar el siguiente modelo ampliado:

Modelo: MCO, usando las observaciones 1-2897 (n = 2408)
Variable dependiente: llabpro

	Coeficiente	Desv. Típica	Estadístico *t*	valor *p*	
const	11,2183	0,0661558	169,6	<0,0001	***
age	−0,0128235	0,00205792	−6,231	<0,0001	***
exporter	0,526498	0,0787440	6,686	<0,0001	***
exper	0,00444900	0,00272845	1,631	0,1031	

	Coeficiente	Desv. Típica	Estadístico *t*	valor *p*	
foreign	0,167975	0,102540	1,638	0,1015	
size_cat_1	0,263253	0,0644453	4,085	<0,0001	***
size_cat_2	0,239197	0,0833670	2,869	0,0042	***

Media de la vble. dep.	11,31505	D.T. de la vble. dep.	1,418927
Suma de cuad. residuos	4596,945	D.T. de la regresión	1,383689
R-cuadrado	0,051422	R-cuadrado corregido	0,049052
F(6, 2401)	21,69299	Valor p (de F)	5,89e-25
Log-verosimilitud	−4195,305	Criterio de Akaike	8404,610
Criterio de Schwarz	8445,115	Crit. de Hannan-Quinn	8419,343

c) ¿Por qué se han creado dos variables ficticias en lugar de tres? Interprete los coeficientes de las variables _cat_1 y Dsize_cat_2.

d) Indique cuál puede ser la razón por la cual las variables *exper* y *foreign* dejan de ser estadísticamente significativas en el modelo anterior.

Ejercicio 5.4

Busque en la red una muestra de datos de *corte transversal* para tres variables continuas de índole económico-empresarial que crea que puedan estar relacionadas.[17] Utilizando dicha muestra de datos, realice las siguientes tareas:

a) Utilizando Excel, guarde en columnas las cuatro variables, nombrándolas y ordenándolas, junto a una variable índice i = 1, 2, ... , N para representar la dimensión de corte transversal. No olvide indicar la fuente de donde se han obtenidos los datos, así como el significado de cada variable y sus unidades de medida. Adicionalmente, construya una cuarta variable ficticia que tome el valor 1 o 0, según si los individuos ide la muestra pertenecen o no a cierta categoría (e.g., género, raza, nacionalidad, continente, sector, países desarrollados, idioma, etc.)

b) Estime por MCO e interprete la correspondiente FRM basada en un modelo tipo:

17. Posibles fuentes de datos: Gapminder (www.gapminder.org), Goolzoom (www.goolzoom.es), Instituto Nacional de Estadística (www.ine.es), Eurostat (https://ec.europa.eu/eurostat/data/database), OECD (https://stats.oecd.org/), World Bank (https://data.worldbank.org/), UNCTADSTAT (https://unctadstat.unctad.org/), FAOSTAT (http://www.fao.org/faostat/en/#data), otros (https://www.economicsnetwork.ac.uk/links/sources, https://db.nomics.world/).

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \gamma_0 D + u \qquad \text{(b1)},$$

donde Y representa la variable dependiente, X_1 y X_2 las variables explicativas continuas, y D una variable explicativa ficticia. Utilice la teoría económica o el razonamiento lógico para justificar qué variable es la dependiente y cuál/es la/s explicativa/s.

c) Basándose en el modelo (b1) anteriormente propuesto, plantee, explique y realice un contraste de significatividad global de la regresión.
d) Basándose en el modelo (b1) anteriormente propuesto, plantee, explique y realice un contraste de Chow.
e) En base a las variables de las que se dispone, plantee y estime un modelo que le permita contrastar si el efecto marginal de X_2 sobre Y depende o no de la pertenencia de i a una de las categorías recogidas por la variable ficticia D.

e1. Plantee y realice dicho contraste.
e2. Represente gráficamente la relación entre Y y X_2, según el modelo estimado.

Referencias

Bibliografía básica

Matilla, Mariano, Pedro Pascual y Basilio S. Carnero). *Econometría y predicción*. UNED: McGraw Hill.

Wooldridge, Jeffrey M. 2014. *Introducción a la econometría. Un enfoque moderno*, 5.ª edición. México: Cengage Learning Editores.

Bibliografía complementaria

Baiocchi, Giovanni y Walter Distaso. 2003. «GRETL: Econometric software for the GNU generation». *Journal of Applied Econometrics* 18 (1): 105-110.

Greene, William H. 1999. *Análisis econométrico*. 3.ª edición. McGraw Hill.

Gujarati, Damodar N. y Dawn C. Porter. 2009. *Econometría*. 5.ª edición. McGraw Hill.

Ramanathan, Ramu. 2002. *Introductory Econometrics with Applications*. 5.ª edición. Orlando, FL: Harcourt College Publishers.

Stock, James H. y Mark W. Watson. 2012. *Introduction to Econometrics*. 3.ª edición. Pearson.

Verbeek, Marno. 2008. *A guide to modern Econometrics*. Chichester, RU: John Wiley & Sons.

Fuentes de datos

Banco Mundial (https://data.worldbank.org): Data_covergence.gdt

Banco Mundial, Doing Business (http://www.doingbusiness.org)

CEPII «gravity» (http://www.cepii.fr)

Datos simulados: Data_marks.xsls, Data_marks2.xsls, Data_production.xlsx.

Encuesta anual de Estructura Salarial del Instituto Nacional de Estadística (http://www.ine.es/): Data_salarios2014ESP.gdt

European Social Survey (www.europeansocialsurvey.org)

Eurostat (https://ec.europa.eu/eurostat/data/database),: Data_cons_inc.xlsx

Gapminder (www.gapminder.org): Data_Gapminder_2010.gdt

Geoportal de hidrocarburos (http://geoportalgasolineras.es), Ministerio de Energía, Turismo y Agenda Digital

ISBN: 978-84-17900-23-6

DOI: http://dx.doi.org/10.6035/Sapientia163

Goolzoom (www.goolzoom.es)
Inside Airbnb (http://insideairbnb.com/)
Latin American Migration Project (LAMP), COL14 (año 2009) (https://lamp.opr.princeton.edu/)
Nestoria (https://www.nestoria.es): Data_Valencia_pisos.gdt(15 abril 2018), Data_Palma_Mallorca_alquileres.gdt(27 agosto 2018)
R&D Scoreboard de la Comisión Europea (http://iri.jrc.ec.europa.eu/scoreboard16.html): Data_RD_scoreboard.gdt

Otras fuentes de interés

DB Nomics (https://db.nomics.world/)
Food and Agriculture Organization of the United Nations, FAOSTAT (http://www.fao.org/faostat/en/#data),
Organization for Economic Cooperation and Development, OECD (https://stats.oecd.org/)
The Economics Network (https://www.economicsnetwork.ac.uk/links/sources)
United Nations Conference on Trade and Development, UNCTADSTAT (https://unctadstat.unctad.org/)

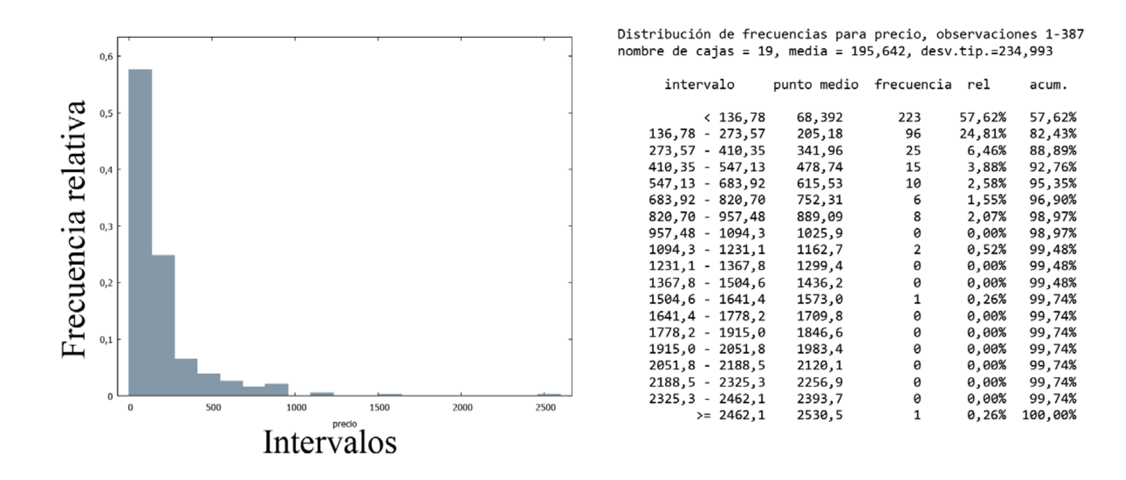
Soluciones de muestra

Solución ejercicio 0.1.

a) Datos de corte transversal, dado que el archivo contiene información para distintas viviendas en un momento del tiempo determinado.

b) Ruta: Click derecho sobre la variable precio / distribución de frecuencias.

Agrupamos los precios de las viviendas en intervalos excluyentes del mismo grosor para, posteriormente, determinar el número relativo de observaciones que hay en cada uno de ellos.



Distribución de frecuencias para precio, observaciones 1-387
nombre de cajas = 19, media = 195,642, desv.tip.=234,993

intervalo	punto medio	frecuencia	rel	acum.
< 136,78	68,392	223	57,62%	57,62%
136,78 - 273,57	205,18	96	24,81%	82,43%
273,57 - 410,35	341,96	25	6,46%	88,89%
410,35 - 547,13	478,74	15	3,88%	92,76%
547,13 - 683,92	615,53	10	2,58%	95,35%
683,92 - 820,70	752,31	6	1,55%	96,90%
820,70 - 957,48	889,09	8	2,07%	98,97%
957,48 - 1094,3	1025,9	0	0,00%	98,97%
1094,3 - 1231,1	1162,7	2	0,52%	99,48%
1231,1 - 1367,8	1299,4	0	0,00%	99,48%
1367,8 - 1504,6	1436,2	0	0,00%	99,48%
1504,6 - 1641,4	1573,0	1	0,26%	99,74%
1641,4 - 1778,2	1709,8	0	0,00%	99,74%
1778,2 - 1915,0	1846,6	0	0,00%	99,74%
1915,0 - 2051,8	1983,4	0	0,00%	99,74%
2051,8 - 2188,5	2120,1	0	0,00%	99,74%
2188,5 - 2325,3	2256,9	0	0,00%	99,74%
2325,3 - 2462,1	2393,7	0	0,00%	99,74%
>= 2462,1	2530,5	1	0,26%	100,00%

Por defecto, los datos se han agrupado en 19 intervalos, que es el número próximo a n, donde n es el número de viviendas (387). El punto central del primer y último intervalo se corresponden, respectivamente, con los valores mínimo y máximo de la muestra de precios.

- Casi el 60 % de los pisos de la muestra tienen un precio de venta < 136.780 €.
- Casi un 25 % de los pisos de la muestra tienen un precio ≥ 136.780 € y < 273.570 €.
- Solamente un piso tiene un precio ≥ 2.462.100 €.

c) Los estadísticos descriptivos ayudan a retratar tres propiedades importantes de un conjunto de datos: la posición, la dispersión y la forma de su distribución.

ISBN: 978-84-17900-23-6

DOI: http://dx.doi.org/10.6035/Sapientia163

Ruta: Click derecho sobre la variable precio / estadísticos principales.

Medidas de posición

Media: $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{x_1+x_2+\cdots+x_n}{n}$

$\overline{precio} = 195.640$ €

$\overline{m2} = 118{,}59\ m^2$

Mediana: Valor central de la distribución de datos ordenados.

$Me(precio_i) = 123.000$ €

$Me(m2_i) = 100\ m^2$

Medidas de dispersión

Rango o recorrido: $Max - Min$

$rango(precio_i) = 2.462.100$ €

$rango(m2_i) = 881\ m^2$

Desviación típica: $\widehat{sd}(x) = s_x = \sqrt{\frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{n-1}}$

$s_{precio} = 234.990$ €

$s_{m2} = 80{,}119\ \ m^2$

Cuanto más concentrados están los datos alrededor de $\bar{x}$, entonces el rango(x_i) y s_x estarán más próximos a 0. En estos casos, las medidas de posición serán más representativas del conjunto de observaciones.

Inconveniente. El rango y la desviación típica dependen de las unidades de medida de la variable analizada, lo cual dificulta la comparación de la representatividad de dos conjuntos de datos expresados en unidades distintas.

Solución. Coeficiente de variación (CV): $\frac{s_x}{|\bar{x}|}$ si $\bar{x} \neq 0$

$V_{precio} = 1{,}201 > 1 \rightarrow s_{precio} > \overline{precio}$ La media es poco representativa del conjunto de datos.

$CV_{m2} = 0{,}676 < 1 \rightarrow s_{m2} < \overline{m2}$ La media es representativa del conjunto de datos.

Medidas de forma

Coeficiente de asimetría

$$CA = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^3}{\left(\sqrt{\frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{n}}\right)^3}$$

Si CA = 0 → $\bar{x} = Me(x)$ Datos distribuidos simétricamente alrededor $\bar{x}$.

Si CA > 0 → $\bar{x} > Me(x)$ La cola derecha de la distrib. es más larga.

Si CA < 0 → $\bar{x} < Me(x)$ La cola izquierda de la distrib. es más larga.

$CA(precio_i) = 4{,}207 > 0$ y $CA(m2_i) = 6{,}116 > 0$

El exceso de curtosis mide la mayor o menor concentración de datos alrededor de $\overline{x}$, versus alrededor de las colas. El nivel de referencia es el correspondiente a una distribución normal (3).

$$EC = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^4}{\left(\sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n}}\right)^4} - 3$$

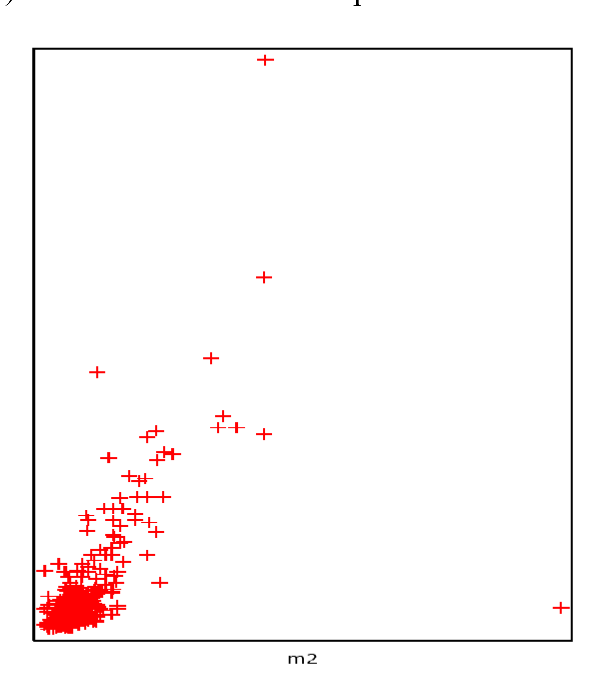
Si EC = 0 Dist. normal

Si EC > 0 Dist. leptocúrtica

Si EC < 0 Dist. platicúrtica

$EC(precio_i) = 27{,}583 > 0$ y $EC(m2_i) = 53{,}296 > 0$

d) Ruta: Ver / Gráficos múltiples / Gráficos X-Y (scatters)



Los pisos más grandes (pequeños) se corresponden mayoritariamente con aquellos pisos más caros (baratos).

e) Ruta: Selecciona variables de interés / click derecho / Matriz de correlaciones

m2	dormitorios	precio	
1,0000	0,6608	0,5534	**m2**
	1,0000	0,4476	**dormitorios**
		1,0000	**precio**

La tabla está mostrando el coeficiente de correlación muestral de Pearson entre cada pareja de variables consideradas:

ISBN: 978-84-17900-23-6

DOI: http://dx.doi.org/10.6035/Sapientia163

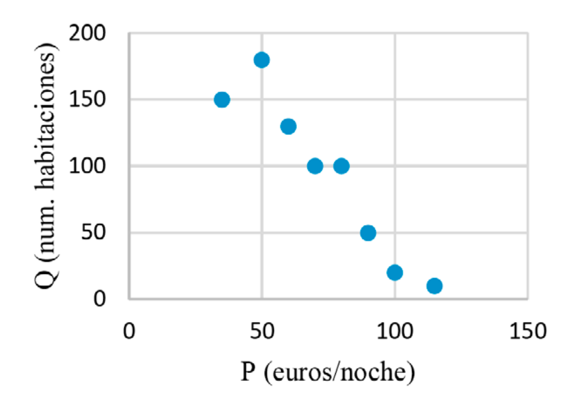
$$r_{xy} = \frac{s_{xy}}{s_x s_y} = \frac{\frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{n-1}}{\sqrt{\frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{n-1}}\sqrt{\frac{\sum_{i=1}^{n}(y_i-\bar{y})^2}{n-1}}} \quad \text{donde} \quad -1 \leq r_{xy} \leq 1$$

- Si $r_{xy}=1 \rightarrow$ Relación lineal positiva perfecta entre *x* e *y*.
- Si $r_{xy}=0 \rightarrow$ No relación lineal entre *x* e *y*, aunque puede existir una relación no lineal.
- Si $r_{xy}=-1 \rightarrow$ Relación lineal inversa perfecta entre *x* e *y*.

Tal y como podemos ver en la matriz de correlaciones, se aprecia una relación lineal positiva relativamente fuerte entre los m2 y el precio ($r_{m2\ precio} = 0{,}553 > 0{,}5$). En cambio, dormitorios y precio presentan una relación lineal positiva, relativamente débil ($r_{dormitorios\ precio} = 0{,}448 < 0{,}5$).

Solución ejercicio 1A.1.

a) Ruta Excel: Insertar / Gráfico X Y (Dispersión)



Se aprecia una relación inversa entre el precio y el número de habitaciones ocupadas. Aquellos hoteles con precios más altos (bajos) ocupan menos (más) habitaciones.

b)

	x_i	y_i	$(x_i-\bar{x})$	$(y_i-\bar{y})$	$(\bar{x}_i-\bar{x})(y_i-\bar{y})$	$(x_i-\bar{x})^2$
hotel (id)	P_i	Q_i	$(P_i-\bar{P})$	$(Q_i-\bar{Q})$	$(P_i-\bar{P})(Q_i-\bar{Q})$	$(P_i-\bar{P})^2$
1	35	150	-40	57,5	-2300	1600
2	100	20	25	-72,5	-1812,5	625
3	90	50	15	-42,5	-637,5	225
4	115	10	40	-82,5	-3300	1600

	x_i	y_i	$(x_i-\bar{x})$	$(y_i-\bar{y})$	$(x_i-\bar{x})(y_i-\bar{y})$	$(x_i-\bar{x})^2$
hotel (id)	P_i	Q_i	$(P_i-\bar{P})$	$(Q_i-\bar{Q})$	$(P_i-\bar{P})(Q_i-\bar{Q})$	$(P_i-\bar{P})^2$
5	70	100	-5	7.5	-37.5	25
6	60	130	-15	37.5	-562.5	225
7	50	180	-25	87.5	-2187.5	625
8	80	100	5	7.5	37.5	25

$\bar{P}$ = 75	$\bar{Q}$ = 93

$\sum_{i=1}^{n}(P_i-\bar{P})(Q_i-\bar{Q})$ = -10800	$\sum_{i=1}^{n}(P_i-\bar{P})^2$ = 4950

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(P_i-\bar{P})(Q_i-\bar{Q})}{\sum_{i=1}^{n}(P_i-\bar{P})^2} = \frac{-10800}{4950} = -2{,}1818$$

$$\hat{\beta}_0 = \bar{Q} - \hat{\beta}_1\bar{P} = 93 - (-2{,}1818)75 = 256{,}1364$$

Función de regresión muestral: $\hat{Q} = 256{,}1364 - 2{,}1818\,P$

Si el precio por noche fuese de 0, se estima que serían ocupadas 256 habitaciones.

Ante un aumento del precio de un 1 €/noche, se estima que el número de habitaciones ocupadas disminuiría en 2,1818 unidades.

c)

ΔP	$\Delta\hat{Q}$
1 €/noche	- 2,1818 habitaciones ocupadas
10 €/noche	- 21,818 habitaciones ocupadas

d)

	x_i	y_i	Valores ajustados $\hat{y}_i$	Residuos $\hat{u}_i=(y_i-\hat{y}_i)$
hotel (id)	P_i	Q_i	$\hat{Q}i$	
1	35	150	$\hat{Q}_1$=256,1364+(-2,1818)·35=179,773	-29,773
2	100	20	$\hat{Q}_2$=256,1364+(-2,1818)·100=37,955	-17,955
3	90	50	$\hat{Q}_3$=256,1364+(-2,1818)·90=59,773	-9,773

	x_i	y_i	**Valores ajustados** $\hat{y}_i$	**Residuos** $\hat{u}_i=(y_i-\hat{y}_i)$
hotel (id)	P_i	Q_i	$\hat{Q}i$	
4	115	10	$\hat{Q}_4$=256,1364+(-2,1818)·115=5,227	4,773
5	70	100	$\hat{Q}_5$=256,1364+(-2,1818)·70=103,409	-3,409
6	60	80	$\hat{Q}_6$=256,1364+(-2,1818)·60=125,227	4,773
7	50	200	$\hat{Q}_7$=256,1364+(-2,1818)·50=147,045	32,955
8	80	100	$\hat{Q}_8$=256,1364+(-2,1818)·80=81,591	18,409

e) Coeficiente de determinación:

$$R^2 = 1 - \frac{SCE}{STC} \qquad donde \qquad SCE = \sum_{i=1}^{n} \hat{u}_i^2 \qquad y \qquad STC = \sum_{i=1}^{n}(y_i - \bar{y})^2$$

	x_i	y_i	**Residuos al cuadrado**	$(y_i-\bar{y})^2$
hotel (id)	P_i	Q_i	$\hat{u}_i^2$	$(Q_i-\bar{Q})^2$
1	35	150	-29,7732 = 886,415	3.306,25
2	100	20	-17,9552 = 322,366	5.256,25
3	90	50	-9,7732 = 95,506	1.806,25
4	115	10	4,7732 = 22,779	6.806,25
5	70	100	-3,4092 = 11,622	56,25
6	60	80	4,7732 = 22,779	1.406,25
7	50	200	32,9552 = 1.086,002	7.656,25
8	80	100	18,4092 = 338,895	56,25
			$\sum_{i=1}^{n} \hat{u}_i^2$ = 2786,364	$\sum_{i=1}^{n}(y_i - \bar{y})^2$ = 26350

Entonces, $R^2 = 1 - \frac{2786,364}{26350} = 0,8943$

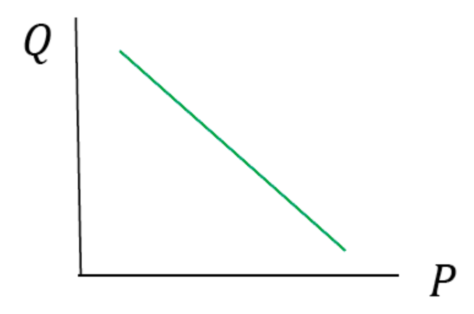
El 89,43 % de la variabilidad muestral exhibida por la demanda es explicada por el precio.

f) $\hat{Q}$ = 256,1364 - 2,1818 75 = 92,5 habitaciones ocupadas

g) Elasticidad precio de la demanda estimada para un precio de 75 euros/noche:

$$\hat{\epsilon}_p^d = \frac{\Delta\hat{Q}}{\Delta P} \cdot \frac{P}{\hat{Q}} = -2{,}1818 \cdot \frac{75}{92{,}5} = -1{,}769$$

Para un precio de 75 euros/noche se estima que, ante un aumento del 1 % del precio, el número de habitaciones ocupadas disminuirá un 1,769 %.



h) Modelo log-log: $\log(Q_i) = \beta_0 + \beta_1 \log(P_i) + u$

En este caso, el parámetro β_1 nos indicaría directamente la elasticidad precio de la demanda (constante). El modelo representaría una función de demanda isoelástica; es decir, con una elasticidad constante con independencia del nivel de precios:



Solución ejercicio 2A.1

a) $y = \beta_0 + \beta_1 x + u$	Modelo lineal en parámetros y en las variables
b) $log\ (y) = \beta_0 + \beta_1\ x + u$	Modelo lineal en parámetros, no lineal en variables
~~*c*) $y = \beta_0 + \sqrt{\beta_1}\ x + u$~~	~~Modelo lineal en las variables, no lineal en parámetros~~
d) $y = e^{\beta_0} x^{\beta_1} e^u$	Tras transformación logarítmica, modelo lineal en parámetros pero no lineal en variables: $\log(y) = \beta_0 + \beta_1 \log(x) + u$
e) $y = \beta_1 + \beta_2 x + \beta_3 x^2 + u$	Modelo lineal en parámetros, no lineal en las variables
f) $y = \beta_1 + \beta_2\ (\frac{1}{x}) + u$	Modelo lineal en parámetros, no lineal en las variables